



"Ss. Cyril and Methodius" University in Skopje
FACULTY OF COMPUTER
SCIENCE AND ENGINEERING

Proceedings of the 22nd International Conference on Informatics and Information Technologies
CiiT 2025

April 25–27, 2025
Hotel Sirius, Strumica, R. N. Macedonia



ISBN 978-608-4699-22-4

Skopje, 2025

22nd International Conference on Informatics and Information Technologies

April 25–27, 2025

Hotel Sirius – Strumica, North Macedonia

Organized by





"Ss. Cyril and Methodius" University in Skopje

FACULTY OF COMPUTER SCIENCE AND ENGINEERING

International Conference on Informatics and Information Technologies

Website: https://ciit.finki.ukim.mk
Email: ciit@finki.ukim.mk

Publisher:

Faculty of Computer Science and Engineering, Skopje, N. Macedonia,
Ss. Cyril and Methodius University in Skopje, N. Macedonia
Address: Rugjer Boshkovikj 16, P.O. Box 393, 1000 Skopje, N. Macedonia

Website: http://www.finki.ukim.mk/
Email: contact@finki.ukim.mk

Proceedings Editors: Boban Joksimoski and Emil Stankov
Technical editing: Boban Joksimoski and Emil Stankov

ISBN: 978-608-4699-22-4

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

004.7:621.39(062)
004(062)

INTERNATIONAL Conference on Informatics and Information Technologies (22 ; 2025 ; Strumica) 22nd International Conference on Informatics and Information Technologies: CiiT 2025, April 25–27, 2025, Hotel Sirius – Strumica, North Macedonia [електронски извор]/ [editors Boban Joksimoski, Emil Stankov]. - Skopje : Faculty of Computer Science and Engineering, 2025

Начин на пристапување (URL):
http://ciit.finki.ukim.mk/proceedings/ciit-2025-proceedings.pdf

- Текст во PDF формат, содржи 379 стр., – Наслов преземен од екранот.
- Опис на изворот на ден 15.12.2025. – Библиографија кон трудовите

ISBN 978-608-4699-22-4

а) Информациско-комуникациски технологии – Собири
б) Компјутерски науки – Собири

COBISS.MK-ID 67716357

Preface

This volume contains the papers presented at the 22nd International Conference on Informatics and Information Technology (CiiT 2025), held from April 25 to 27, 2025, at Hotel Sirius in Strumica, North Macedonia. The conference was organized by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Republic of North Macedonia.

Over the past twenty-two years, CiiT has established itself as a prominent regional and international platform for the dissemination of research and innovation in the fields of informatics and information technologies. The conference continues to provide an inspiring environment for researchers, professionals, and students to exchange ideas, present original research, and discuss emerging trends shaping the digital landscape.

Building upon the success of previous editions, the 2025 conference featured a diverse selection of papers and presentations covering various domains of computer science and IT. As in tradition, special sessions were dedicated to the work of talented undergraduate and master students, offering them a stage to present their research results and practical projects. Outstanding student contributions were recognized with awards for the best paper and best presentation.

This year, we were honored to host three distinguished invited speakers whose expertise enriched the conference program. Prof. Matjaz Gams, Ph.D., from the Department of Intelligent Systems at the Jožef Stefan Institute, Ljubljana, Slovenia, delivered a keynote lecture titled "From LLMs to Superintelligence." Prof. Dr. Reiner Creutzburg, Professor of Applied Computer Science at TH Brandenburg and Professor for IT Security at SRH Berlin, Germany, presented "IT Forensics – An Overview." Prof. Stevo Lukić, MD, PhD, Associate Professor at the Department of Neurology, Medical Faculty, University of Niš, and the Clinic of Neurology, University Clinical Center Niš, Serbia, delivered a talk on "Human Evaluation Framework within the ChatMed Project."

The success of CiiT 2025 was made possible through the continued support of our academic and industrial partners, as well as the enthusiasm and dedication of all contributors and participants.

We believe that this year's conference once again served as a catalyst for innovation, collaboration, and the exchange of transformative ideas, reinforcing the mission of CiiT to foster progress and excellence in informatics and information technology.

We would like to express our sincere gratitude to all authors for their valuable contributions, and for sharing their insights, research, and expertise. Your work plays a vital role in advancing the field of informatics and information technology and in shaping the future of digital innovation.

On behalf of the organizing committee, we extend our heartfelt thanks to all participants, reviewers, and supporters of CiiT 2025. Your engagement and collaboration have been essential to the success of this year's conference. We hope that CiiT 2025 has provided you with new knowledge, inspiration, and opportunities for future cooperation within the vibrant and evolving community of information technology researchers and practitioners.

May, 2025
Skopje, N. Macedonia

Prof. Dr. Boban Joksimoski
Prof. Dr. Emil Stankov

Organization

Conference Chairs

Boban Joksimoski, Ph.D.	Assistant Professor – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Emil Stankov, Ph.D.	Assistant Professor – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Metodija Jancheski, Ph.D.	Associate Professor – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia

Organizing Committee

Mile Jovanov, Ph.D.	Professor – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Stefan Andonov, M.Sc.	Teaching/Research Assistant – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Petar Sekuloski, M.Sc.	Teaching/Research Assistant – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Vlatko Spasev, M.Sc.	Teaching/Research Assistant – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Zhivko Atanaskoski, B.Sc.	Demonstrator – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Dimitar Mileski, M.Sc.	Demonstrator – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Mila Dodevska, B.Sc.	Demonstrator – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Marija Taneska, B.Sc.	Demonstrator – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia

Slave Temkov, B.Sc.	Demonstrator – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Marko Petrov, B.Sc.	Demonstrator – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Ema Pandilova, B.Sc.	Demonstrator – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Elena Atanasoska, B.Sc.	Demonstrator – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
Marija Stojcheva, B.Sc.	Demonstrator – Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia

Program Committee

Nevena Ackovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Vangel Ajanovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ljupcho Antovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Goce Armenski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Hrachya Astsatryan	Institute for Informatics and Automation Problems, National Academy of Sciences of Armenia, Armenia
Amelia Badica	University of Craiova, Romania
Tsonka Baicheva	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Verica Bakeva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Antun Balaz	Institute of Physics, Belgrade, Serbia

Lasko Basnarkov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Galina Bogdanova	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Adrijan Božinovski	Faculty of Informatics at the American University of Europe, Skopje, N. Macedonia
Özge Büyükdağlı	International University of Sarajevo, Bosnia and Herzegovina
Eliot Bytyçi	University of Prishtina, Prishtina, Kosovo
Alessandro Cantelli-Forti	Lab RaSS National Laboratory - CNIT, Pisa, Italy
Ivan Chorbev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Betim Cico	EPOKA University, Tirana
Aleksandra Dedinec	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Konstantin Delchev	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Boris Delibasic	Faculty of Organizational Sciences, University of Belgrade, Serbia
Vesna Dimitrievska	Silicon Austria Labs, Villach, Austria
Vesna Dimitrievska Ristovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Vesna Dimitrova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ivica Dimitrovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Tome Eftimov	Computer Systems, Jožef Stefan Institute, Slovenia
Ilche Georgievski	University of Stuttgart, Germany
Sonja Gievska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Dejan Gjorgjevikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Danilo Gligoroski	Department of Information Security and Communication Technology, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Norway
Sasho Gramatikov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Marjan Gusev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Norbert Hounkonnou	University of Abomey-Calavi, Benin
Natasha Ilievska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Bojan Ilijoski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ilinka Ivanoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Edmond Jajaga	Mother Teresa University, Skopje, N. Macedonia
Boro Jakimovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Metodija Jancheski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Smilka Janeska-Sarkanjac	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Valentina Janev	Institute Mihajlo Pupin, Belgrade, Serbia
Boban Joksimoski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Mile Jovanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Milos Jovanovik	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Vacius Jusas	Kaunas University of Technology, Lithuania
Slobodan Kalajdziski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Stoyan Kapralov	Technical University - Gabrovo, Bulgaria
Ivan Kitanovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Magdalena Kostoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Bojana Koteska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Andrea Kulakov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Arianit Kurti	Linnaeus University, Växjö, Sweden
Petre Lameski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Igor Ljubi	University of Zagreb, Croatia
Suzana Loshkovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ana Madevska Bogdanova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Gjorgji Madjarov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Smile Markovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Hristina Mihajloska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Marija Mihova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Aleksandra Mileva	Faculty of Computer Science, University Goce Delcev, Shtip, N. Macedonia
Georgina Mirceva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Miroslav Mirchev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Kostadin Mishev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Igor Mishkovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Kosta Mitreski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Melanija Mitrovic	University of Niš, Serbia
Tarik Namas	International University of Sarajevo, Bosnia and Herzegovina
Andreja Naumoski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Marcin Paprzycki	Systems Research Institute, Polish Academy of Sciences, Poland
Dilip Patel	London South Bank University, London, United Kingdom
Ivan Miguel Pires	Instituto de Telecomunicações, ESTGA, Universidade de Aveiro, Portugal
Nuno Pombo	University of Beira Interior, Portugal
Aleksandra Popovska-Mitrovikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Panche Ribarski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Blagoj Ristevski	University St. Kliment Ohridski, Faculty of Information and Communication Technologies, Bitola, N. Macedonia
Simona Samardjiska	Digital Security Group, Radboud University, The Netherlands
Snezana Savoska	University St. Kliment Ohridski, Faculty of Information and Communication Technologies, Bitola, N. Macedonia
Loren Schwiebert	Wayne State University, Detroit, MI, USA
Sahra Sedigh Sarvestani	Missouri University of Science and Technology, Rolla, Missouri, USA
Josep Silva	DSIC, Universidad Politécnica de Valencia, Spain
Monika Simjanoska Misheva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Petar Sokoloski	Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Marko Spasenovic	Institute of Chemistry, Technology and Metallurgy, Belgrade, Serbia
Dejan Spasov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Emil Stankov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Eugenia Stoimenova	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Riste Stojanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Biljana Stojkoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Natasha Stojkovikj	University Goce Delcev, Shtip, N. Macedonia
Aleksandar Stojmenski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Stanimir Stoyanov	University of Plovdiv "Paisii Hilendarski Bulgaria
Biljana Tojtovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Dimitar Trajanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ljiljana Trajkovic	Simon Fraser University, Burnaby, Canada
Vladimir Trajkovik	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Kire Trivodaliev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Katarina Trojacanec-Dineva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Goran Velinov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Dusan Vudragovic	Institute of Physics Belgrade, Belgrade, Serbia

Massimiliano Zanin IFISC (CSIC-UIB), Palma de Mallorca, Spain

Vladimir Zdraveski Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Eftim Zdravevski Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Katerina Zdravkova Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Stela Zhelezova Institute of Mathematics and Informatics, BAS, Bulgaria

Table of Contents

1 **Embedded Device for Efficient Watering of Agricultural Land**
Kristijan Gjoshev, Biljana Risteska Stojkoska 1

2 **Detection of Atrial Fibrillation: A Systematic Review and Meta-Analysis**
Sara Gjorgjieva, Riste Stojanov, Marjan Gusev 5

3 **Food Data in the Semantic Web: A Review of Nutritional Resources, Knowledge Graphs, and Emerging Applications**
Darko Sasanski, Riste Stojanov 11

4 **Robotics, Electrical Engineering, and Tourism in Macedonia**
Stevo Bozinovski, Vangel Bozinoski 17

5 **Optimization of the Race Strategy in Formula 1**
Marija Saveska, Dimitar Trajanov 21

6 **AI-Powered Pipeline for Video Generation from Large Texts**
Andrej Nankov, Vladimir Trajkovik 27

7 **Topologically Equivalent Binary Trees Regarding Binary Tree Roll**
Adrijan Božinovski 33

8 **Implementing Interactive Learning For Educational Involvement Via Algorithmic Game Design**
Marko Angelovski, Bojan Zivkovikj, Andon Angov, Viktor Denkovski, Marija Stankova Medarovska 38

9 **A Custom Mocking Server for API Simulation: Design, and Implementation**
Rilind Kaba, Fisnik Doko, Martin Djonov, Stojan Kitanov 42

10 **Graph-Based Approaches to Input-Output Analysis, Trade Flows, and Economic Complexity: A Literature Review**
Dimitar Peshevski, Dimitar Trajanov 48

11 **VulnerSec: A Flexible, Automated and Open-Source Cybersecurity Framework**
Evgenija Krajchevska, Vojdan Kjorveziroski 54

12 **Platform for Environmental Data Crowdsensing and Visualization**
Pance Cavkovski, Slave Temkov, Aleksandar Nikov, Petre Lameski, Vladimir Trajkovik 60

13 **Chronic Kidney Disease Detection Using Mass Spectrometry Data: A Survey**
Hana Hasanicaj, Riste Stojanov 66

14 **Escape to Learn: Assessing the Role of Game-Based Learning in Cybersecurity Education**
Maja Videnovik, Vladimir Trajkovik 73

15 **End-to-End Control of a F1Tenth Vehicle**
Atanasko Boris Mitrev, Andrea Kulakov 78
16 **Evaluating Large Language Models Potential in Mental Health Counseling: A Comparative Study Using the Therapy Empathy Scale**
Borjan Georgievski, Niki Ramancev, Marija Stankova Medarovska 82
17 **Hyperparameter Optimization and AutoML of a Simple CNN for Skin Lesion Classification**
Marino Jakimoski, Ilinka Ivanoska 87
18 **Autism Spectrum Disorder Classification Using Functional Brain Connectivity**
Jonatan Mileski, Ilinka Ivanoska 91
19 **Performance Analysis of Open-Source Large Language Models: Accuracy, Efficiency, and Response Variability**
Vesna Pop-Dimitrijoska Koteska, Ana Markovska, Andrej Petkovikj, Marija Maneva, Ema Pandilova, Pance Ribarski, Bojan Ilijoski 95
20 **Comparison of Hénon Map With Logistic Map Performance in Encryption and Decryption of Images**
Zhivko Shishovski, Aleksandar Stojmenski, Ivan Chorbev 101
21 **AI-Driven Project Management: LLM-Powered Task Automation and Optimization**
Ilija Bozoski, Sasho Gramatikov 105
22 **Data Preprocessing and Exploration of Developmental Disabilities Dataset**
Ana Markovska, Bojan Ilijoski 109
23 **AI Based Cyberbullying Detection - State of the Art**
Mahnoor Jamil, Ivan Chorbev, Boban Joksimoski 114
24 **Large Language Models for Legal Reasoning in Low Resource Languages**
Ivan Apostolovski, Martin Angelovski, Vladimir Radevski, Marija Stankova Medarovska 118
25 **Bitcoin Price Prediction Using Classical Machine Learning and Deep Learning**
Martina Jovanovikj, Hristijan Gjoreski 123
26 **Quantitative Analysis of Tunnel Stability Using a Data-Driven Approach**
Mite Mazgaliev, Stefan Andonov, Bojan Ilijoski, Zlatko Zafirovski 129
27 **Automatic Speech Recognition for Swiss German: A Comparative Analysis of Whisper and Wav2Vec**
Marija Stojcheva, Goran Petkovski, Igor Mishkovski 134
28 **Analysis of a Stochastic Predator-Prey Model for Detective-Criminal Dynamics**
Viktor Najdovski, Verica Bakeva 140
29 **Web-Based Platform for Enhancing Coordination in Technology Transfer**
Marija Karajanova, Dimitar Kitanovski, Zorica Karapancheva, Aleksandar Stojmenski, Petre Lameski, Ivan Chorbev 144

30 **Modelling the Dynamics of Heterogenous Populations**
Jana Hristova, Verica Bakeva 149
31 **Exploration into LLMs Capabilities for Question-answering over Tabular Data**
Teodora Mladenovska, Martina Toshevska, Sonja Gievska 153
32 **Web Tracking Technologies - Current Research Analysis**
Resul Bedii Gümüş, Boban Joksimoski 157
33 **Python for Mass Spectrometry: A Comprehensive Overview of Libraries for Data Processing and Analysis**
Elena Atanasoska, Riste Stojanov 162
34 **Overview of Foundation Models in Remote Sensing**
Teo Stojkovski, Evrosina Stojkoska, Marko Petrov, Ema Pandilova, Vlatko Spasev, Ivan Kitanovski 168
35 **Accelerating IoV Intrusion Detection: Benchmarking GPU-Accelerated vs CPU-Based ML Libraries**
Furkan Colhak, Hasan Coşkun, Tsafac Nkombong Regine Cyrille, Tedi Hoxa, Mert Ilhan Ecevit, Mehmet Nafiz Aydın 174
36 **Analysis on Features of Human Glaucoma using Human Glaucoma Dataset**
Sara Trpchevikj, Petar Sekuloski 180
37 **Interactive Recommendation System with Graph Neural Networks**
David Arsovski, Dimitar Trajanov, Marija Saveska 186
38 **Android Digital Forensics of Vulnerable Mobile Applications: Data Extraction and Analysis of Hardcoded Strings**
Isiaq Bolaji Ibrahim 192
39 **Analysis and Comparison of Super-Resolution Models**
Boris Manev, Dime Gjorgjiev, Sasho Gramatikov 198
40 **Navigating the Data Space Landscape: Concepts, Applications, and Future Directions**
Bojana Marojevikj, Riste Stojanov 202
41 **Time-Series Heart Rate Variability with CUDA**
Andrej Shekerov, Dimitar Mileski, Marjan Gusev 208
42 **Farm and Animal Intelligent Recommendation System - FAIRS**
Zoran Petrovski, Ana Madevska Bogdanova 212
43 **Enhancing AI Education with a Custom Code Runner Template: A Solution for Fair Grading and Improved Feedback in Search-Based Problems**
Martin Dinev 217
44 **Computational Photography Assessment: A Framework for Image Quality and Composition Analysis**
Valentin Cvetanoski, Ljubinka Sandjakoska, Rasim Salkoski 221
45 **Ontology-Based Structuring and Analysis of North Macedonian Public Procurement Contracts**
Bojan Ristov, Stefan Eftimov, Milena Trajanoska, Dimitar Trajanov . . . 227
46 **AI-Driven Vulnerability Detection in Code: A State-of-the-Art Review with a Focus on CWE and Java Frameworks**
Klesida Gjana 231

47 **AI Incorporated Job Matching System for Creating Career Opportunities**
Narcis Nacev, Nikola Kostadinov, Viktor Denkovski, Marija Stankova Medarovska 235
48 **Ciphertext-Only Attacks Using Large Language Models**
Marija Taneska, Ikechukwu John Chukwu 240
49 **Implementation of Non-Interactive Zero-Knowledge Proof Based Authentication Protocol for Student Authentication**
Zivko Atanaskoski, Zorica Karapancheva, Vesna Dimitrova, Riste Stojanov 245
50 **Online Shoppers Purchasing Intention Prediction and Analysis Using Machine Learning**
Natalija Arizankoska, Hristijan Gjoreski 250
51 **A Multi-Agent System for Student Banking and Stock Market Analysis**
Darko Trajanov 256
52 **Going NUTS with ADVI: Exploring Various Bayesian Inference Techniques with Facebook Prophet**
Jovan Krajevski, Biljana Tojtovska Ribarski 260
53 **Neural Radiance Fields (NeRFs) for Satellite Imagery: A Literature Review**
Arnen Telkjiu, Dejan Gjorgjevikj 266
54 **Software Implementations of Lightweight Ciphers: A Survey with Focus on ASCON**
Ikechukwu John Chukwu, Vesna Dimitrova 272
55 **Raising Awareness in Protecting from RFID Attacks**
Andrej Vuchkovikj, Sanja Apostolovikj, Nikola Kostadinov, Viktor Denkovski, Marija Stankova Medarovska, Goce Gavrilov 278
56 **Moodle-GPT Chatbot: An AI Moodle Integration for Collaborative Learning**
Nikola Kostadinov, Narcis Nacev, Ekaterina Cvetkova, Andrej Vuchkovikj, Marija Stankova Medarovska, Viktor Denkovski 283
57 **Exploring Transformer Models for EMG Signal Classification: A Case Study in ALS Detection**
Mihail Gjorgjiev, Kostadin Mishev, Monika Simjanoska Misheva 288
58 **Mitigating Malicious Spoofing and Sniffing Attacks Through a Portable Secure Router**
Andrej Vuchkovikj, Viktor Danev, Viktor Denkovski, Bogdan Jeliskoski, Goce Gavrilov 294
59 **Ollama LLMs Serverless Deployment - Performance Evaluation using LangChain**
Zorica Karapancheva, Jovan Simonoski, Marjan Gusev, Dimitar Mileski 299
60 **Combination of Kolb's Experiential Learning Cycle and Project-Based Learning in Data Science Education**
Nevena Kimova, Ana Madevska Bogdanova 305
61 **Digital Forensics in Cloud Computing: Leveraging Google Cloud Forensics Utils**
Isiaq Bolaji Ibrahim 309

62 **Game Theory to Cryptography by Incentivizing Secure Federated Learning in Cyber Contexts**
Simon Atanasovski, Elissa Mollakuqe, Vesna Dimitrova 314
63 **A Survey on Intelligent Agentic Workflows**
Agon Osmani, Riste Stojanov 321
64 **A Review of Cross-Layer Security Vulnerabilities in IoT Communication Protocols**
Berfin Ebrar Atabey, Sasho Gramatikov 327
65 **A State-of-the-Art Review on Social Media Exposure and Cyber Threats**
Lydiah Wanjugu Muigwa, Ivan Chorbev 331
66 **Large Language Models for Predictive Neurology: A Neurogenomic and Neurogenetic Perspective**
Djansel Bukovec, Blagojche Gjorgjioski, Kostadin Mishev, Monika Simjanoska Misheva 335
67 **Mapping the Landscape of Anonymous CTI Sharing**
Asem Mousa, Petre Lameski 340
68 **Hate Speech Detection on Social Media Using Transformer-Based Models: A BERT Approach**
Mihail Trajkovikj, Marko Lazarevski, Anabela Bochvarova, Evgenij Anushev, Ilija Trajkovski, Georgina Mirceva 345
69 **Access Control Using Threshold Cryptography and Dynamic Policy Management**
Aldiyar Ismailov 349
70 **Evaluating Speech-to-Text Models for Macedonian and English: A Comparative Study**
Cvetanka Nechevska, Ivan Kitanovski 354
71 **Leveraging Topological Data Analysis for Enhanced S&P 500 Price Forecasting**
Petar Sekuloski, Dobrinka Sotirova 359
72 **Using the BBC Micro:bit in Educational Settings: Recommendations for N. Macedonia**
Gorica Miceva, Mile Jovanov, Emil Stankov, Zivko Atanaskoski 363
73 **Exploring the Educational Potential of Virtual Reality and Mixed Reality: Immersive Learning, Student Engagement, and Knowledge Retention**
Mila Dodevska, Zivko Atanaskoski, Mile Jovanov, Emil Stankov, Petre Lameski 367
74 **Confidence in the Mirror, Doubt in the Crowd: Navigating Trust in the Age of AI**
Tanja Pavleska, Vladimir Trajkovikj 373

Embedded Device for Efficient Watering of Agricultural Land

Kristijan Gjoshev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
gjoshev.kristijan@students.finki.ukim.mk

Biljana Risteska Stojkoska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
biljana.stojkoska@finki.ukim.mk

***Abstract*—This research is targeting a development of an embedded device that will be used for efficient watering of agricultural land. The device will be realized through use of hardware based on ESP32 microcontroller and a software solution based on Arduino development environment. The focus was on making the device easy to use, economically viable, and great attention was paid to fault tolerance and error handling since the device is meant to be used without supervision for extended periods of time. The device was designed with the latest technological advancements in mind and built in a way that is easily upgradable and adaptable to the location where it will be used.**

***Index Terms*—water, plants, agriculture, land, automation, efficiency, ESP32, Arduino**

I. Introduction

Water is one of the basic elements of nature, necessary for existence of the living world. The need for fresh and clean water is growing, and growing plants requires a considerable amount of water. The rational usage of this resource becomes a key part of successful agriculture and social responsibility. As every plant has water needs as well as tolerance for lack thereof, so growing plants in relatively dry conditions and watering as needed is the easiest way to control the watering process. Since this watering process is done manually by humans through some sort of water delivery method there's a chance to automate it and improve it. The industry already has automation for this process, but that automation is not adequate for small producers and people that produce just for themselves. To fill that gap the aim is to create a device that will be cheap, easy to use and will be adequate for these user groups.

II. Watering as Part of the Process for Growing Plants

A. The Importance of Water

The water has two objectives in the life-cycle of plants, one is as part of the photosynthesis process, and another one is as a transport media for nutrients from the soil. The current agricultural practices use water as a transport media for the chemistry used in treatment of plants and land. Better control of the watering process allows to reduce the amounts of chemicals needed which in turn gives better soil and plant quality.

B. Socially Responsible Use of Water

While it's ideal to use non-potable water for plants, it's not always possible to do that. Using potable water to do watering is sometimes the only option, so rational usage of it is a must to ensure there's enough water for basic human needs in the area during periods of drought.

C. Room for Improvement of the Watering Process

The motivation for this research as a real life problem on a location in the Ohrid region. The region is known for wast amounts of water available, but also very dry periods. There are big temperature swings during the day as well as rainfalls that do not correlate to the weather forecasts. On this particular piece of land there's a diversity of plants that all share the same soil. There are some food producing plants and decorative flowers that require watering directly at the root, they can't survive well without water for more than 2-3 days. There's also grass and green surfaces that require water showers instead.

D. Requirements

According to the problem we are facing the following requirements were taken into consideration:

1) The device needs to be able to control multiple water flows, because of the different delivery methods
2) The device needs to be able to react on the current state of the soil because of the unpredictable weather conditions
3) The device needs to be able to perform activities according to a time schedule
4) There must be provided means for manual control and adjustments on the spot
5) The device must not be too expensive compared to the savings it brings

Before building this device a market research was done for existing similar devices. Almost all devices can do activities based on a time schedule, and they could all be controlled on their permanent installation location. However, not a single device was able to work with soil moisture input. Devices that could control more than one flow, such as the commercially available Rain Bird SST1200OUT [1], cost more than 200€ in a complete setup.

III. Hardware

A. Choice of microcontroller

The choice of microcontroller to handle all input and output came down to ESP32 [2] by Esspressif. The development board (Fig. 1) of choice can be powered by 5V, and the microprocessor itself works with 3.3V logic. To keep track of time DS1302 real-time clock was used. The user preferences will be stored on a MicroSD card. The user input will be done through a rotary encoder, and to provide user feedback a monochromatic display was chosen based on the SSD1306 controller. The bill of materials is 37€ total.

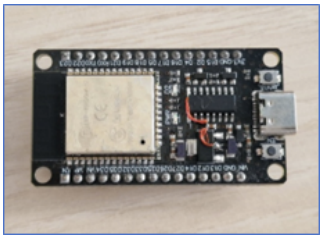


Fig. 1. ESP32 development board

B. Solenoid Driver

The water flow is controlled through solenoid valves which run on 12V. To be able to switch the solenoids a circuit was designed to utilize the 3.3V logic from the microcontroller. The solenoids have resistance of 24 Ohms, so according to Ohm's law there will be current of 0.5A at 12V operating voltage. To deal with this current and voltage adequate MOSFET semiconductor was needed. The MOSFET of choice was IRL540N [3] by Infineon, mostly because it can be driven directly with 3.3V and still provide 20A of current - more than enough. The supporting elements are a diode (D1) that is present for protection of the MOSFET from the voltage spikes that happen when switching an inductor, operation indicating diode D1 and its current limiting resistor R2, and the R1 resistor that brings the MOSFET to a stable OFF state when there's no input. The solenoid driver schematics can be seen in Fig. 2, and the complete motherboard with four solenoid drivers on top can be seen in Fig 3.

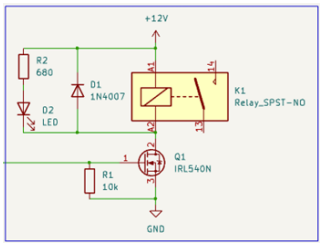


Fig. 2. MOSFET based solenoid switching building block

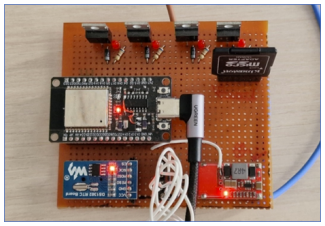


Fig. 3. The motherboard of the device

C. Soil Moisture Sensing

An attempt was made for resistive sensing of moisture content in the soil by using the water and minerals in the soil as conductor. Copper probes were used to measure the resistance of the soil across distance of 5cm. However the measurements across time were not consistent, and the idea was abandoned until better, more consistent way of measuring soil moisture content is found.

IV. Software

The software for the microcontroller was done in a Arduino development environment [4] and was build with object-oriented paradigms in mind.

A. Initialization

The default state of the microcontroller has all pins set as inputs. This needs to be corrected according to the connections of the peripheral hardware. After the pins are ready the peripheral hardware is initialized as well according to the manufacturers recommendations and needed adjustments. For this purpose the "setup()" function is given, it is run right after boot. The initialization logic is ordered in a way that all dependencies are loaded sequentially. First the most basic elements are loaded that have no prerequisites, such as the encoder, the display, the clock, and an empty menu. The user configuration is then loaded, and is supplied to the menu and clock libraries.

B. Repeatable Tasks

After the initialization is done the device starts with its routine tasks. That includes dealing with the user input (buttons and encoder motion separately), checks for possible reset trigger conditions, processes menu actions according to the user input, does adjustments to the display according to the changes in the menu, adjusts the config to reflect the menu actions, and finally processes the timers and relays.

C. Synchronizing the Clock and Timed Events

For communication with the DS1302 RTC chip the library RTC [5] by Makuna, available on Github. Since communication with the RTC chip is relatively slow the clock is read on boot, the ESP32 internal clock is synced to it and further time queries are done against the ESP32 internal clock. This library also provides alarm logic that covers the needs for timed events. Two alarms are needed for every watering event

- start and finish, so for 4 solenoids each having 5 timed event there are 40 alarms in use.

D. User Interface

A monochromatic display based on the SDD1306 chip was used to provide user feedback. The library U8g2lib [6] by olikraus was found to be adequeate for interfacing with it. The library supports various graphic drawing capabilities for basic geometric shapes, as well as text support. In combination with the library LinkedList [7] by ivanseidel a menu interface was made with multiple levels and dynamic contents. The menu types available include boolean toggle, integer adjustment, date, time and list pick.

E. User Input

For user input a rotary encoder with a button was used. Both of those inputs have imperfections and noise in their output, so some input sanitization was needed. The rotary encoder is handled through a finite state machine [8] which deals with any jumps there may be. The button was debounced in software by waiting for stable state of at least 50ms before confirming the state change.

Because some of the values on the device have pretty wide range the rotational acceleration of the rotary encoder was taken into account as a indicator of user intention for faster adjustment. Depending on the acceleration a jump of 10, 100 or 500 values per tick was given.

The change of state on the rotary encoder can happen very fast, so the encoder is handled through the use of processor interrupts. The intention of change is registered, then the processor implements the change of state in a repeatable task later on.

F. Configuration Persistence

All configured values are kept in a JSON object managed by the ArduinoJSON [9] library. The object is then serialized and written to a MicroSD card. For any value not present in the JSON a safe, default value is used.

Operations on the JSON object outside of the device can be done as an alternative way of configuring the device. It's also the way to make a backup of the configuration.

The device can be reset by removing this JSON object from the MicroSD card, effectively loading the default values on boot.

The device can also function without saving the configuration to a MicroSD card, however the persistance across reboots will be lost.

G. Control of the Solenoids

The solenoid valves as an electro-mechanical devices have their own inertia, static friction and coil saturation, which can be used for more efficient control over them. The solenoids produce 6W of heat per hour at 12V and 0.5A operation. This voltage and current are only needed to move the solenoid. After the movement is completed, which takes no more than 200ms, there is static friction helping the valve to stay in position. This allows to reduce the voltage supplied to the solenoid and still keep it open, and by that the current and heat dissipation are also lowered. This value was empirically estimated to be 7.2V, which is 60% of the operational voltage. To achieve this voltage the output signal driving the MOSFETs was adjusted to Pulse-Width modulated signal with 60% duty cycle. After the voltage is reduced the current through the solenoid is only 0.3A, which translates to 2.16W of heat per hour.

The switching of the PWM signal needs to be fast enough for the solenoid to stay saturated during the Off period of the duty cycle. The choice of frequency was 1KHz, and was decided as a middle ground across three states:

1) Audible humming in the range between 4KHz and 20KHz
2) Instability of the saturation in the solenoid at frequencies bellow 50Hz
3) Reduced efficiency of the MOSFET because of too high switching frequencies above 50KHz

H. Responsiveness of the device

The Arduino development environment for ESP32 microcontrollers uses both cores of the microcontroller. The development environment is based on FreeRTOS, operating system suitable for managing time-critical processes. By design, one of the cores is dedicated to dealing with Wifi, Bluetooth, system tasks, housekeeping and watchdogs. The other core is left for user activities. While both cores can be used for user activities upon a design decision, it was decided against that in the name of device stability. By using one core the risks of potential concurrent access and resource starvation are avoided.

This leads to another case where all functions need to be non-blocking, or with blocking short enough to be unnoticeable by the user. The maximum blocking for watering events is 1 second, since all processing of timers is done on 1 second intervals. The maximum blocking duration for display refresh is 33.3 milliseconds, as the refresh rate of the display is set to be 30Hz. This is the interval taken as maximum time for reflecting the user input on the display. To fulfill this 33.3ms requirement some routines are based on triggers, where a trigger is set when something needs to be accomplished and the activity is completed through one or more runs of the loop() section.

V. Real world testing

A. Hardware

A small watering setup was made to test the hardware. 10L reservoir was used as a source of water with 4 solenoid valves used for water distribution. Another alternative setup was made with an electric water pump instead one of the valves and 3 valves were used for water distribution. The water pump was run directly in place of a solenoid valve, with the PWM duty-cycle adjusted to 100%, effectively making that output constant 12V for the pump.

B. Software

It was discovered that deactivating and reactivating a watering event caused corruption of the watering event alarms. Further analysis showed that the RTC library had use-after-free bug that was triggered at object creation through the copy-constructor and assignment operator. The bug was fixed and was reported to the author of the library.

VI. Fault tolerance

Greater attention to detail was given to fault tolerance since this device will need to work without supervision for extended periods of time. Some of the faults are mitigated through hardware design, the rest are mitigated through software.

A. Power supply interruptions

It's considered acceptable that the device can be inoperable when there's a power outage, but the device must go in a safe state if power is removed at any moment. The solenoid valves are normally-closed, so they fall shut when power is removed. The electronics are designed in a way that they fall back to default state during the boot process, so that power outage recoveries are handled in a stable manner.

B. Time de-synchronization

This fault state would mean that the device will be operating at wrong time, breaking one of the key requirements. There's a battery for the time keeping module so that it keeps track during power outages. When that battery dies the clock resets to a time in the past - a de-synchronization occurs. This reset to a time in the past is used as an indicator of de-synchronization when compared to the compile time of the software. At that moment the correct time is not known, but it is known that the time is not correct, so all operations are stopped until the user sets the time again.

C. Persistent memory corruption

Two possible fault states are identified - total corruption of the configuration file on the MicroSD card and value corruption while having a valid file. In case of total corruption the safe default settings are loaded and the configuration is rewritten upon first configuration adjustment. In case of value corruption the safe default is used for that particular field only but all activities are halted until the configuration is corrected.

D. Mechanical failures of the valves

This fault was encountered during the testing phase. The solution is to use valve or pump at the input as a master control for the input flow, and one or more valves on output. This way failure of a single valve, especially failing stuck open, will still be a safe failure.

VII. Potential improvements

A. Using external weather forecast services

Since the microcontroller already supports WiFi connections this seems like a logical improvement to do. This would be a major improvement in the economics of water usage by opportunistic delaying of watering events in case of expected rainfall in the very near future.

B. Tracking the local weather

This would include sensors for doing the weather forecasting locally, as well as verifying the external weather forecast services. This is a must-have for correct opportunistic delaying of watering events, as failure of the forecast would mean potential under-watering or over-watering.

C. Remote control and mesh networks

The presence of WiFi, Bluetooth and Bluetooth LE allows for implementation of remote control and communication of the device with other devices and services. This would enable synchronized control, data sharing, telemetry and more.

VIII. Conclusion

With practical use of technology a manual watering process was successfully automated. The device was tested for a short period, during that period improved watering consistency and reduced human errors were observed. With adequate design decisions the fault tolerance is high enough to allow for unsupervised usage of the device. Longer period of testing and observation would be needed to give quantitative measurements of any improvements it may have brought, however using the device to offset the efforts for manual watering is already showing benefits.

The device is affordable enough to motivate test usage, and is built with components that are readily available and easily obtainable. It's also easily upgradable and adaptable to various usage scenarios, and it's easy to adjust and configure on the spot.

Improvements are to be made in the foreseeable future with better sensing technologies and better hardware and software.

The code for the project can be found on Github at https://github.com/Crypter/RelayController.

References

[1] Rain Bird SST1200OUT automatic irrigation device, https://store.rainbird.com/sst1200out-12-station-outdoor-sst-simple-to-set-irrigation-timer.html [Accessed 10.10.2024]
[2] ESP32 by Espressif, https://www.espressif.com/en/products/socs/esp32 [Accessed 10.10.2024]
[3] IRL540N specifications, https://www.infineon.com/dgdl/Infineon-IRL540N-DataSheet-v01_01-EN.pdf [Accessed 28.11.2024]
[4] Arduino environment for ESP32, https://github.com/espressif/arduino-esp32 [Accessed 12.10.2024]
[5] RTC Library by Makuna, https://github.com/Makuna/Rtc [Accessed 17.10.2024]
[6] U8g2 monochromatic display library by olikraus, https://github.com/olikraus/u8g2 [Accessed 17.10.2024]
[7] LinkedList library by ivanseidel, https://github.com/ivanseidel/LinkedList [Accessed 12.12.2024]
[8] Rotary encoder as finite state machines documentation, https://www.bristolwatch.com/arduino/arduino2.htm [Accessed 16.11.2024]
[9] JSON library for Arduino, https://arduinojson.org/ [Accessed 12.12.2024]

Detection of Atrial Fibrillation: A Systematic Review and Meta-Analysis

Sara Gjorgjieva
Innovation Dooel
Skopje, North Macedonia
sara.gjorgjieva@innovation.com.mk

Riste Stojanov
Faculty of Computer Science and Engineering
Skopje, North Macedonia
riste.stojanov@finki.ukim.mk

Marjan Gusev
Faculty of Computer Science and Engineering
Skopje, North Macedonia
marjan.gushev@finki.ukim.mk

***Abstract*—Early detection of heart conditions is crucial in preventing severe complications and saving lives. Among them, atrial fibrillation (AF) is a common cardiac arrhythmia that significantly increases the risk of stroke and heart failure if left undiagnosed. However, AF detection remains challenging due to its intermittent nature and variability in clinical presentation. This paper reviews machine learning (ML) and deep learning (DL) methods for early AF detection, exploring traditional models such as Decision Trees (DT), Random Forests (RF), and XGBoost (XGB), as well as advanced techniques like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Siamese Neural Networks (SNN). Additionally, it highlights key aspects such as feature extraction, appropriate evaluation metrics, and challenges related to data imbalance and the integration of multiple data sources. By addressing these factors, this review underscores the potential of ML and DL in improving the accuracy and efficiency of AF detection.**

***Index Terms*—Atrial fibrillation, AF, ECG, machine learning, deep learning**

I. Introduction

Atrial fibrillation (AF) is a common cardiac arrhythmia where the heart's rhythm becomes irregular and rapid due to uncoordinated quivering of the atria, which impacts heart function [1]. AF is classified into three types: paroxysmal (episodes last minutes to a week and resolve on their own), persistent (lasting more than seven days and may require treatment), and permanent (continuous irregular rhythm without attempts to restore normalcy) [2].

Symptoms of AF vary widely; some individuals experience none, while others may suffer from shortness of breath, dizziness, or fatigue. Despite the absence of symptoms, AF significantly raises the risk of stroke due to the formation of blood clots in the heart. Early detection is crucial, as highlighted by studies linking silent AF episodes to increased stroke risk, emphasizing the need for effective screening strategies [3].

Diagnosis typically relies on ECG to detect irregular heart rhythms. ECG technology and machine learning (ML) have improved the accuracy of AF detection. The main idea behind using ML for AF detection is that these algorithms can learn patterns in the data that may not be obvious to the human eye [4]. Studies have demonstrated that ML models can accurately identify AF by leveraging large datasets, outperforming traditional diagnostic methods in specific scenarios [5]. By training on these datasets, ML models can become very accurate at identifying whether a person has AF.

With the increasing availability of ECG data and ML techniques, AF detection has become a growing field of study. However, the task is complex due to the variability of AF presentation, differences in datasets, and challenges related to data quality and imbalance.

This review aims to provide an AF-specific perspective on ML-based detection by examining datasets used, tasks defined within the literature, and the models developed to solve those tasks. We also discuss feature engineering and key challenges specific to AF detection.

II. Datasets for AF detection

The foundation of AF detection lies in the ECG datasets used for model training and evaluation. Table I summarizes the most common datasets used in AF research. These datasets vary significantly in terms of signal length, number of subjects, sampling frequency, number of leads, and annotation granularity. For example, some datasets such as AFDB and LTAFDB [6] provide long-term Holter recordings with detailed AF episode annotations, while others like the PhysioNet CinC 2017 dataset consist of shorter, single-lead recordings with categorical rhythm labels. Moreover, datasets such as MITDB [7] and UCI Arrhythmia offer a broader range of arrhythmia types, supporting multiclass classification tasks beyond binary AF detection. The diversity in dataset characteristics plays a crucial role in determining model robustness, generalizability across patient populations, and real-world clinical applicability. This datasets can be found on the Physionet site [8].

III. Models and Methods for AF detection

ML models play a crucial role in detecting AF. Even though challenging, with their help, the detection of this cardiac arrhythmia has become much more manageable. However, some models can achieve better results; we analyzed the

TABLE I
OVERVIEW OF PUBLIC DATASETS USED FOR AF DETECTION

Dataset	Source	Size	Notes
AFDB	PhysioNet	25 subjects	Annotated AF episodes
LTAFDB	PhysioNet	84 records	Long-term ECG recordings with AF labels
NSRDB	PhysioNet	18 records	Normal sinus rhythm (NSR)
IAFDB	PhysioNet	23 records	AF and other arrhythmias
MITDB	PhysioNet	48 half-hour recordings	Multiple arrhythmia types including AF
PhysioNet CinC 2017	PhysioNet Challenge	8,528 single-lead ECGs	AF, normal, other rhythms, noise
PhysioNet CinC 2020	PhysioNet Challenge	12-lead ECGs from 43,101 patients	Includes 27 different cardiac diagnoses
UCI Arrhythmia Dataset	UCI ML Repository	452 instances, 279 features	16-class arrhythmia classification

most common ML models that eventually led to the best results. The most popular supervised ML models used for detecting AF are Decision Trees (DTs), Random Forest (RF), XGBoost, AdaBoost, etc. Regarding the DL models, the RNN, DNN, and SiameseNN play a significant role in detecting this heart problem. AF-specific learning tasks go beyond generic classification. Based on the literature that we reviewed, the major ML tasks in AF include: Each of these tasks has specific data, model, and evaluation requirements.

A. AF Classification

Atrial Fibrillation classification is one of the most commonly addressed tasks in AF-related ML research. It involves distinguishing between AF and non-AF segments in ECG recordings. This task has been tackled using both classical machine learning models and deep learning architectures.

1) Tree-Based and Ensemble Methods: Classical models such as Decision Trees (DT), Random Forests (RF), AdaBoost, and XGBoost have demonstrated competitive performance for AF classification.

Decision Trees (DT) are among the earliest and most interpretable models used for AF classification.These models classify ECG segments by leveraging features such as RR intervals, P-wave morphology, and heart rate variability (HRV), using a hierarchical structure to identify patterns indicative of AF episodes. While Decision Trees are efficient and interpretable, they can be prone to overfitting, especially with high-dimensional data, necessitating techniques like feature selection and pruning to improve performance [9]. For instance, DT-based thresholding techniques achieved F1-score exceeding 97% on datasets like LTAFDB and AFDB.

Random Forests (RF) enhance the robustness of Decision Trees by averaging multiple trees trained on bootstrapped samples.RF is highly effective in AF detection, leveraging its ability to handle imbalanced datasets and extract meaningful insights from diverse ECG features, such as spectral entropy, HRV metrics, and wavelet transform coefficients. The model's inherent feature-ranking capability aids in identifying the most critical predictors of AF, making it a valuable tool for feature selection in clinical studies [10]. A study using proprietary data reported an F1-score of 95.56% using RF [11].

Boosting methods, such as XGBoost and AdaBoost, are widely used in AF detection for their ability to enhance model performance by combining multiple weaker classifiers. XGBoost is particularly effective in capturing complex nonlinear relationships within ECG features, such as morphological changes and frequency-domain characteristics, often outperforming traditional ML approaches [12], [13]. AdaBoost, on the other hand, excels in improving the accuracy of simpler base classifiers, such as Decision Trees, by focusing on features like RR intervals, P-wave duration, and heart rate dynamics [14] [15], [16]. These models have also shown strong results in AF detection tasks. For example, AdaBoost achieved an F1-score of 96.99% on the PhysioNet/CinC 2017 dataset [17], while XGBoost performed similarly well on a combination of several AF databases (F1 = 98.97%) [18].

2) Artificial Neural Networks: Feedforward artificial neural networks (ANNs) were also explored in early works on AF classification. These models performed well on standard ECG datasets such as MITDB and AFDB. In one study, ANN achieved a high F1-score of 98.97% [19]. Similarly, Artis et al. [20] used ANNs and obtained an F1 score of 92.60%

3) Deep Learning Approaches: DL uses models with multiple layers to learn complex patterns in data [21]. These models are effective for detecting AF because they can process large datasets and automatically extract features without manual engineering [22]. The Neural networks, including Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Siamese Neural Networks (SNNs), have emerged as powerful tools in AF detection due to their ability to model complex patterns in ECG data. RNNs, including Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), are well-suited for handling sequential ECG data, capturing temporal dependencies crucial for AF detection [23], [24]. For instance, Faust et al. used an RNN-LSTM hybrid model to achieve 98.35% F1-score using 10-fold cross-validation on AFDB [25]. SNNs, on the other hand, focus on comparing ECG signal segments, proving useful in one-shot learning scenarios where limited labeled data is available [26], [27].

4) Convolutional Neural Networks: CNNs have become a key method in AF detection due to their capability to automatically extract spatial and temporal features from raw or minimally preprocessed ECG data. These networks identify waveform characteristics, such as P-waves and QRS complexes, enabling accurate differentiation between AF and normal rhythms. CNNs are particularly effective for multi-lead ECGs, as they can capture inter-lead relationships to enhance feature extraction. Their real-time applicability in wearable devices further underscores their utility in continuous patient monitoring [28], [29]. Sudies have applied CNNs to PhysioNet/CinC 2017 and MITDB datasets, achieving F1-scores ranging from 71% to 89.9%. For instance, Luo et al. [30] utilized CNNs with 5-fold cross-validation, reaching an F1 score of 83.80%. Cao et al. [31] applied CNNs with 6-fold cross-validation, obtaining an F1 score of 89.90% on the MITDB dataset. Andreotti et al. [32] aimed to improve patient management and reduce healthcare costs by classifying short ECG segments into four categories (AF, normal, other rhythms, or noise) using a CNN, achieving an F1 score of 83% with 5-fold cross-validation. Ghiasi et al. [33] proposed an automatic classification algorithm using CNNs for ECG

waveform analysis, achieving an F1 score of 71%. Maknickas et al. [34] combined CNNs with LSTM to detect AF directly from QRS complex features, achieving an F1 score of 75% on the PhysioNet Challenge 2017 dataset with an 80/20 splitting ratio.

CRNNs and CNN-LSTM hybrid models combine temporal and spatial modeling, yielding robust performance on challenging datasets. Limam et al. [35] proposed a convolutional recurrent neural network (CRNN) for AF detection, achieving an F1 score of 77.00% on the MITDB database with an 85/15 splitting ratio. Wang et al. [36] implemented a hybrid CNN-LSTM framework, achieving an F1 score of 82% on the MIT-BIH Arrhythmia PhysioNet dataset with a 70/30 splitting ratio.

These studies highlight the versatility of CNNs and their variants, including hybrid models, in capturing key ECG signal features, thereby enabling accurate AF detection. Despite challenges such as computational demands and the need for diverse datasets, CNNs remain a powerful tool in automated arrhythmia diagnosis.

B. Multiclass Classification

Multiclass classification expands the binary AF vs non-AF task by including additional rhythm classes, such as normal sinus rhythm (NSR), other arrhythmias (O), and sometimes noisy or artifacted segments. This is especially relevant in real-world datasets like PhysioNet/CinC 2017, which defines four classes: Normal, AF, Other rhythm, and Noisy.

In clinical practice, patients often present with a variety of arrhythmic events, not just AF, making multiclass classification a more realistic and practical modeling task. It allows systems to not only detect AF but also distinguish it from other abnormal rhythms like premature atrial contractions, ventricular tachycardia, or atrial flutter, which may have different prognostic and therapeutic implications. Multiclass approaches typically require more complex modeling strategies and often utilize feature extraction techniques or deep learning architectures that can capture temporal and morphological variations between classes.

C. Classical Machine Learning Approaches

Several classical ML models have been adapted for multiclass classification in AF studies: decision trees, random forests, support vector machines (SVM), have been widely used due to their interpretability and relatively low computational cost. These models often rely on hand-crafted features extracted from ECG signals, such as RR intervals, wavelet coefficients, or statistical descriptors.

Random Forests (RF), Decision Trees (DT) and SVMs were trained on engineered features such as RR intervals, entropy measures, and signal morphology. These models showed good accuracy but limited generalization in highly imbalanced settings. Mustaqeem et al. [37] applied SVM-based multiclass classification with wrapper-based feature selection on a 16-class arrhythmia dataset. The one-against-one SVM approach performed best, achieving 92.07% accuracy with a 90/10 data split. When it comes to using RF as a model for AF detection, the authors proposed a novel arrhythmia classification method using ECG signals, combining wavelet-based denoising, fused local/global features, and PCA for dimensionality reduction. A Random Forest classifier trained on MIT-BIH data achieved 99.08% accuracy intra-patient and 92.31% inter-patient, outperforming SVM and BP neural networks.

Gradient Boosting Models, particularly XGBoost, have also been applied to multiclass ECG classification. [13] On PhysioNet/CinC 2017, XGBoost reached 81% F1-score, performing competitively in the PhysioNet 2017 Challenge, classifying the ECG signals into normal, AF, and other rhythms.

1) Deep Learning Methods: Deep learning models often outperform classical ML on multiclass tasks due to their ability to learn end-to-end features directly from ECG signals. Convolutional Neural Networks (CNNs), in particular, have shown strong performance by effectively capturing local temporal and morphological patterns in ECG data relevant for rhythm classification. CNNs are frequently used for spatial pattern recognition across multiple rhythm types. Liu et al. used deep convolutional neural networks (DCNNs) to classify six arrhythmia types using 10-second photoplethysmography (PPG) signals. Their model achieved 85% accuracy and a microaverage AUC of 0.978, showing the potential of PPG-based deep learning for arrhythmia screening. A spatio-temporal attention CNN-RNN (STA-CRNN) model was proposed from [38] to classify nine ECG rhythm types. The model achieved an F1-score of 0.835 and highlighted clinically relevant features through attention visualization.

IV. Feature Extraction Techniques

One of the critical steps in using ML for AF detection is feature extraction. Features are the essential pieces of information that the model uses to make predictions. Incorporating relevant features directly influences the accuracy and reliability of the model, making this step crucial in the overall pipeline. In the context of ECG data, features might include measurements like the time between heartbeats (RR intervals) [39], the shape of the ECG waves [40], or even more complex mathematical properties of the signal [41].

A. RR Intervals and dBPM Signals

RR intervals refer to the time between successive R-waves on the ECG, which correspond to the peaks of the QRS complex, representing the heartbeat. These intervals are crucial for detecting AF because irregularities in the timing of the beats are a key indicator of the condition [42]. By analyzing RR intervals, clinicians and researchers can gain insights into the heart's rhythm and identify potential episodes of AF [25]. A standard formula calculates beats per Minute (BPM) values from RR intervals. The dBPM values, which represent the difference between two consecutive BPM values, are derived from these BPM calculations. This method is detailed in [13] and was used in the following study [43]. dBPM signals highlight the changes in heart rate over time, providing another valuable feature for detecting AF. This derivative measure enhances the model's ability to detect subtle variations in heart

rate that might not be immediately apparent from the raw ECG data alone.

B. Entropy-Based Features

Entropy-based features, derived from information theory, are valuable in binary classification models for AF detection as they quantify the complexity or unpredictability of ECG signals. Shannon entropy measures the randomness in ECG signals, with higher values often indicating irregular heart rhythms associated with AF [44]. Its application in ECG analysis has been shown to enhance the detection of irregular patterns characteristic of AF [45]. Similarly, sample entropy focuses on signal complexity, making it particularly effective for distinguishing AF from other arrhythmias, as highlighted in multiple studies [46] [47], [48]. Another innovative feature is derived using GZIP compression algorithms, where the compression ratio reflects signal irregularity; more complex signals yield lower compression efficiency, offering a novel approach to AF detection that has shown promising results [49], [50]. These features can be conveniently extracted using tools like the TSFEL (Time Series Feature Extraction Library). This simplifies analyzing time series data, enabling researchers to focus on model development and optimization [51]. Integrating these entropy-based features enhances model performance, contributing to more accurate AF detection.

V. Evaluation Metrics

Evaluating the performance of models for AF detection requires robust metrics to ensure reliability and effectiveness. Although widely recognized, accuracy can be misleading in imbalanced datasets like those used for AF detection, as it often favors the majority class [52]. The F1 score, the harmonic mean of precision and recall, provides a more balanced evaluation by capturing the trade-off between false positives and false negatives [53], [54]. Precision, which measures the proportion of true positives among all positive predictions, is crucial in clinical settings to avoid unnecessary interventions caused by false positives [55]. Conversely, recall (or sensitivity) evaluates the proportion of true positives among all actual positive cases, underscoring the importance of minimizing missed AF diagnoses [56].

When it comes to using multiclass classification for AF detection, typically the macro-averaged F1-score is the most used metric, along with precision and recall to evaluate performance. Due to clinical relevance, recall for the AF class remains the most emphasized metric, even in multiclass settings.

VI. Challenges in using ML for AF detection

Detecting AF using ML faces several challenges, particularly data quality and class imbalance. The quality of ECG signals is crucial, as noisy or artifact-laden recordings can hinder model performance [62], [63]. Preprocessing techniques have been employed to address these issues and enhance data reliability [64], [65]. Another critical challenge is the imbalance in datasets, where AF cases are often underrepresented compared to normal rhythms [28]. This imbalance can lead to biased models that struggle to detect AF cases accurately. To counteract this, methods like data augmentation, oversampling, and cost-sensitive learning have been proposed, helping models better identify minority class cases while maintaining overall performance [66], [67], [68]. Addressing these challenges is vital to ensuring the generalizability of models across diverse patient populations and varying data collection environments [69], [70], [71]. Moreover, improving interpretability remains essential, as healthcare applications require a clear understanding of model predictions to support patient treatment decisions [72].

VII. Conclusion

In this survey, we explored ML and DL methods for detecting AF, covering models from traditional approaches like Decision Trees and XGBoost to advanced techniques such as CNNs, RNNs, and Siamese Neural Networks. The success of these models depends on key factors like effective feature extraction, proper data preprocessing, and suitable evaluation metrics. We also addressed challenges such as poor data quality, class imbalance, and the need for models that can perform well in diverse real-world scenarios, highlighting strategies like data augmentation and cost-sensitive learning to overcome these issues.

Despite significant advancements, further research is needed to improve AF detection by integrating diverse datasets, enhancing real-time performance, and ensuring models are interpretable for clinical use. ML and DL promise to enable earlier and more accurate AF diagnoses. As these technologies evolve, their integration into healthcare systems will significantly improve the quality of the patients' lives.

References

[1] A. L. Waldo, "Mechanisms of atrial fibrillation," *Journal of cardiovascular electrophysiology*, vol. 14, pp. S267–S274, 2003.

[2] J. Pillarisetti, A. Patel, K. Boc, S. Bommana, Y. Sawers, S. Vanga, H. Sayana, W. Chen, J. Nath, J. Vacek, *et al.*, "Evolution of paroxysmal atrial fibrillation to persistent or permanent atrial fibrillation: predictors of progression," *Journal of atrial fibrillation*, vol. 2, no. 1, 2009.

[3] R. A. Bernstein, H. Kamel, C. B. Granger, J. P. Piccini, P. P. Sethi, J. M. Katz, C. A. Vives, P. D. Ziegler, N. C. Franco, L. H. Schwamm, *et al.*, "Effect of long-term continuous cardiac monitoring vs usual care on detection of atrial fibrillation in patients with stroke attributed to large- or small-vessel disease: the stroke-af randomized clinical trial," *Jama*, vol. 325, no. 21, pp. 2169–2177, 2021.

[4] A. K. Feeny, M. K. Chung, A. Madabhushi, Z. I. Attia, M. Cikes, M. Firouznia, P. A. Friedman, M. M. Kalscheur, S. Kapa, S. M. Narayan, *et al.*, "Artificial intelligence and machine learning in arrhythmias and cardiac electrophysiology," *Circulation: Arrhythmia and Electrophysiology*, vol. 13, no. 8, p. e007952, 2020.

[5] A. M. Sanchez de la Nava, F. Atienza, J. Bermejo, and F. Fernandez-Aviles, "Artificial intelligence for a personalized diagnosis and treatment of atrial fibrillation," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 320, no. 4, pp. H1337–H1347, 2021.

[6] S. Petrutiu, A. V. Sahakian, and S. Swiryn, "Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans," *EP Europace*, vol. 9, pp. 466–470, 05 2007.

[7] R. Mark and G. Moody, "Mit-bih arrhythmia database," *See http://ecg. mit. edu/dbinfo. html*, 1997.

[8] "Physionet site, datasets,"

TABLE II
COMPARISON OF AF CLASSIFICATION METHODS (BINARY AND MULTICLASS) WITH EVALUATION METRICS

Paper	Dataset	Method	Split	SEN (%)	PPV (%)	F1 Score (%)
[9]	LTAFDB/AFDB	DT	separate ds	97.37	97.89	97.63
[57]	MITDB	DT	-	97.00	94.00	95.50
[19]	MITDB	ICA	-	100.00	99.72	99.85
[11]	Proprietary	RF	50/50	100.00	91.50	95.56
[18]	NSRDB, AFTDB, AFDB, LTAFDB, IAFDB	XGB	50/50	98.43	99.51	98.97
[17]	PhysioNet/CinC 2017	AdaBoost	-	96.70	98.00	96.99
[19]	MITDB, AFDB	ANN	-	99.60	99.72	98.97
[20]	MITDB	ANN	-	92.86	92.34	92.60
[58]	Proprietary	DDNN	5-fold CV	99.19	-	90.70
[59]	PhysioNet/CinC 2017	RNN	-	-	-	83.00
[25]	AFDB	RNN + LSTM	10-fold CV	98.32	98.39	98.35
[35]	PhysioNet/CinC 2017	CRNN	85/15	72.70	-	-
[31]	PhysioNet/CinC 2017	CNN	6-fold CV	-	97.0	89.9
[32]	PhysioNet/CinC 2017	CNN	5-fold CV	-	-	83.00
[33]	PhysioNet/CinC 2017	CNN	97/3	-	-	71.00
[34]	PhysioNet/CinC 2017	LSTM	80/20	-	-	75.00
[36]	MITDB	CNN+LSTM	70/30	-	-	82.00
[60]	PPG (6-class)	DCNN	-	75.80	75.20	-
[61]	ECG (4-class)	AdaBoost	-	-	-	82.50
[13]	-	Piecewise spline + XGB	10-fold CV	77.90	81.20	81.00
[38]	12-lead ECG (8-class)	STA-CRNN	4-fold CV	-	-	83.50

[9] X. Zhou, H. Ding, W. Wu, and Y. Zhang, "A real-time atrial fibrillation detection algorithm based on the instantaneous state of heart rate," *PloS one*, vol. 10, no. 9, p. e0136544, 2015.

[10] V. Kalidas and L. S. Tamil, "Detection of atrial fibrillation using discrete-state markov models and random forests," *Computers in biology and medicine*, vol. 113, p. 103386, 2019.

[11] S. Mehrang, O. Lahdenoja, M. Kaisti, M. J. Tadi, T. Hurnanen, A. Airola, T. Knuutila, J. Jaakkola, S. Jaakkola, T. Vasankari, *et al.*, "Classification of atrial fibrillation and acute decompensated heart failure using smartphone mechanocardiography: a multilabel learning approach," *IEEE Sensors Journal*, vol. 20, no. 14, pp. 7957–7968, 2020.

[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[13] Y. Chen, X. Wang, Y. Jung, V. Abedi, R. Zand, M. Bikak, and M. Adibuzzaman, "Classification of short single-lead electrocardiograms (ecgs) for atrial fibrillation detection using piecewise linear spline and xgboost," *Physiological measurement*, vol. 39, no. 10, p. 104006, 2018.

[14] R. E. Schapire, "Explaining adaboost," in *Empirical inference: festschrift in honor of vladimir N. Vapnik*, pp. 37–52, Springer, 2013.

[15] M. Shao, G. Bin, S. Wu, G. Bin, J. Huang, and Z. Zhou, "Detection of atrial fibrillation from ecg recordings using decision tree ensemble with multi-level features," *Physiological measurement*, vol. 39, no. 9, p. 094008, 2018.

[16] J. Cao, S. Kwong, and R. Wang, "A noise-detection based adaboost algorithm for mislabeled data," *Pattern recognition*, vol. 45, no. 12, pp. 4451–4465, 2012.

[17] I. S. Mahmood and I. A. M. Abdelrahman, "A comparison between different classifiers for diagnoses of atrial fibrillation," in *2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pp. 1–6, IEEE, 2019.

[18] C. A. Millán, N. A. Girón, and D. M. Lopez, "Analysis of relevant features from photoplethysmographic signals for atrial fibrillation classification," *International Journal of Environmental Research and Public Health*, vol. 17, no. 2, p. 498, 2020.

[19] R. J. Martis, U. R. Acharya, H. Prasad, C. K. Chua, C. M. Lim, and J. S. Suri, "Application of higher order statistics for atrial arrhythmia classification," *Biomedical signal processing and control*, vol. 8, no. 6, pp. 888–900, 2013.

[20] S. G. Artis, R. Mark, and G. Moody, "Detection of atrial fibrillation using artificial neural networks," Master's thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and ..., 1991.

[21] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: a new paradigm to machine learning," *Archives of Computational Methods in Engineering*, vol. 27, pp. 1071–1092, 2020.

[22] S. P. Shashikumar, A. J. Shah, Q. Li, G. D. Clifford, and S. Nemati, "A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology," in *2017 IEEE EMBS international conference on biomedical & health informatics (BHI)*, pp. 141–144, Ieee, 2017.

[23] S. Grossberg, "Recurrent neural networks," *Scholarpedia*, vol. 8, no. 2, p. 1888, 2013.

[24] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[25] O. Faust, A. Shenfield, M. Kareem, T. R. San, H. Fujita, and U. R. Acharya, "Automated detection of atrial fibrillation using long short-term memory network with rr interval signals," *Computers in biology and medicine*, vol. 102, pp. 327–335, 2018.

[26] D. Chicco, "Siamese neural networks: An overview," *Artificial neural networks*, pp. 73–94, 2021.

[27] Y. Ng, M.-T. Liao, T.-L. Chen, C.-K. Lee, C.-Y. Chou, and W. Wang, "Few-shot transfer learning for personalized atrial fibrillation detection using patient-based siamese network with single-lead ecg records," *Artificial Intelligence in Medicine*, vol. 144, p. 102644, 2023.

[28] G. Petmezas, K. Haris, L. Stefanopoulos, V. Kilintzis, A. Tzavelis, J. A. Rogers, A. K. Katsaggelos, and N. Maglaveras, "Automated atrial fibrillation detection using a hybrid cnn-lstm network on imbalanced ecg datasets," *Biomedical Signal Processing and Control*, vol. 63, p. 102194, 2021.

[29] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Convolutional neural networks," in *Multivariate statistical machine learning methods for genomic prediction*, pp. 533–577, Springer, 2022.

[30] J. Luo, C. Fu, M. Bai, and Y. Zhao, "Atrial fibrillation detection with convolutional neural network," in *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*, pp. 94–98, 2018.

[31] X.-C. Cao, B. Yao, and B.-Q. Chen, "Atrial fibrillation detection using an improved multi-scale decomposition enhanced residual convolutional neural network," *IEEE Access*, vol. 7, pp. 89152–89161, 2019.

[32] F. Andreotti, O. Carr, M. A. Pimentel, A. Mahdi, and M. De Vos, "Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ecg," in *2017 Computing in Cardiology (CinC)*, pp. 1–4, IEEE, 2017.

[33] S. Ghiasi, M. Abdollahpur, N. Madani, K. Kiani, and A. Ghaffari, "Atrial fibrillation detection using feature based algorithm and deep convolutional neural network," in *2017 Computing in Cardiology (CinC)*, pp. 1–4, IEEE, 2017.

[34] V. Maknickas and A. Maknickas, "Atrial fibrillation classification using qrs complex features and lstm," in *2017 Computing in Cardiology (CinC)*, pp. 1–4, IEEE, 2017.

[35] M. Limam and F. Precioso, "Atrial fibrillation detection and ecg classification based on convolutional recurrent neural network," in *2017 Computing in Cardiology (CinC)*, pp. 1–4, IEEE, 2017.

[36] J. Wang and W. Li, "Atrial fibrillation detection and ecg classification based on cnn-bilstm," *arXiv preprint arXiv:2011.06187*, 2020.

[37] A. Mustaqeem, S. M. Anwar, and M. Majid, "Multiclass classification of cardiac arrhythmia using improved feature selection and svm invariants," *Computational and mathematical methods in medicine*, vol. 2018, no. 1, p. 7310496, 2018.

[38] J. Zhang, A. Liu, M. Gao, X. Chen, X. Zhang, and X. Chen, "Ecg-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network," *Artificial Intelligence in Medicine*, vol. 106, p. 101856, 2020.

[39] J. Duan, Q. Wang, B. Zhang, C. Liu, C. Li, and L. Wang, "Accurate detection of atrial fibrillation events with rr intervals from ecg signals," *Plos one*, vol. 17, no. 8, p. e0271596, 2022.

[40] J. Rodriguez, A. Goni, and A. Illarramendi, "Real-time classification of ecgs on a pda," *IEEE Transactions on information Technology in biomedicine*, vol. 9, no. 1, pp. 23–34, 2005.

[41] N. Sadr, M. Jayawardhana, T. T. Pham, R. Tang, A. T. Balaei, and P. de Chazal, "A low-complexity algorithm for detection of atrial fibrillation using an ecg," *Physiological measurement*, vol. 39, no. 6, p. 064003, 2018.

[42] J. Lian, L. Wang, and D. Muessig, "A simple method to detect atrial fibrillation using rr intervals," *The American journal of cardiology*, vol. 107, no. 10, pp. 1494–1497, 2011.

[43] M. Gusev, "System and method for monitoring of atrial fibrillation blood glucose and psychophisical condition based on heart rate variability," *Google Patents: WIPO WO2022144571A1*, 2020.

[44] M. Masè, L. Faes, R. Antolini, M. Scaglione, and F. Ravelli, "Quantification of synchronization during atrial fibrillation by shannon entropy: validation in patients and computer model of atrial arrhythmias," *Physiological measurement*, vol. 26, no. 6, p. 911, 2005.

[45] X. Zhou, H. Ding, B. Ung, E. Pickwell-MacPherson, and Y. Zhang, "Automatic online detection of atrial fibrillation based on symbolic dynamics and shannon entropy," *Biomedical engineering online*, vol. 13, pp. 1–18, 2014.

[46] R. Alcaraz and J. J. Rieta, "A review on sample entropy applications for the non-invasive analysis of atrial fibrillation electrocardiograms," *Biomedical Signal Processing and Control*, vol. 5, no. 1, pp. 1–14, 2010.

[47] R. Alcaraz and J. Rieta, "A novel application of sample entropy to the electrocardiogram of atrial fibrillation," *Nonlinear Analysis: Real World Applications*, vol. 11, no. 2, pp. 1026–1035, 2010.

[48] T. Horie, N. Burioka, T. Amisaki, and E. Shimizu, "Sample entropy in electrocardiogram during atrial fibrillation," *Yonago Acta Medica*, vol. 61, no. 1, pp. 049–057, 2018.

[49] T. Meurers, "Implementation and test of an infrastructure for automatic evaluation of electrocardiograms,"

[50] S. Bharadwaj, "Using convolutional neural networks to detect compression algorithms," in *Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2022*, pp. 33–45, Springer, 2022.

[51] A. Al-Mousa, J. Baniissa, T. Hashem, and T. Ibraheem, "Enhanced electrocardiogram machine learning-based classification with emphasis on fusion and unknown heartbeat classes," *Digital Health*, vol. 9, p. 20552076231187608, 2023.

[52] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks.," *Journal of Machine Learning Research*, vol. 10, no. 12, 2009.

[53] R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models," in *Proceedings of the first workshop on evaluation and comparison of NLP systems*, pp. 79–91, 2020.

[54] M. Muntean and F.-D. Militaru, "Metrics for evaluating classification algorithms," in *Education, Research and Business Technologies: Proceedings of 21st International Conference on Informatics in Economy (IE 2022)*, pp. 307–317, Springer, 2023.

[55] R. Ranawana and V. Palade, "Optimized precision-a new measure for classifier performance evaluation," in *2006 IEEE international conference on evolutionary computation*, pp. 2254–2261, IEEE, 2006.

[56] A. Lavie, K. Sagae, and S. Jayaraman, "The significance of recall in automatic metrics for mt evaluation," in *Machine Translation: From Real Users to Research: 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28-October 2, 2004. Proceedings 6*, pp. 134–143, Springer, 2004.

[57] S. Sieciński, P. S. Kostka, and E. J. Tkacz, "Comparison of atrial fibrillation detection performance using decision trees, svm and artificial neural network," in *Information Technology and Systems: Proceedings of ICITS 2019*, pp. 693–701, Springer, 2019.

[58] W. Cai, Y. Chen, J. Guo, B. Han, Y. Shi, L. Ji, J. Wang, G. Zhang, and J. Luo, "Accurate detection of atrial fibrillation from 12-lead ecg using deep neural network," *Computers in biology and medicine*, vol. 116, p. 103378, 2020.

[59] T. Teijeiro, C. A. García, D. Castro, and P. Félix, "Arrhythmia classification from the abductive interpretation of short single-lead ecg records," in *2017 Computing in cardiology (cinc)*, pp. 1–4, IEEE, 2017.

[60] Z. Liu, B. Zhou, Z. Jiang, X. Chen, Y. Li, M. Tang, and F. Miao, "Multiclass arrhythmia detection and classification from photoplethysmography signals using a deep convolutional neural network," *Journal of the American Heart Association*, vol. 11, no. 7, p. e023555, 2022.

[61] A. Mukherjee, A. D. Choudhury, S. Datta, C. Puri, R. Banerjee, R. Singh, A. Ukil, S. Bandyopadhyay, A. Pal, and S. Khandelwal, "Detection of atrial fibrillation and other abnormal rhythms from ecg using a multi-layer classifier architecture," *Physiological measurement*, vol. 40, no. 5, p. 054006, 2019.

[62] L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch, "The effects of data quality on machine learning performance," *arXiv preprint arXiv:2207.14529*, 2022.

[63] U. Satija, B. Ramkumar, and M. S. Manikandan, "A review of signal processing techniques for electrocardiogram signal quality assessment," *IEEE reviews in biomedical engineering*, vol. 11, pp. 36–52, 2018.

[64] V. Gupta, M. Mittal, and V. Mittal, "A simplistic and novel technique for ecg signal pre-processing," *IETE Journal of Research*, vol. 70, no. 1, pp. 815–826, 2024.

[65] Q. Li, C. Liu, J. Oster, and G. D. Clifford, "Signal processing and feature selection preprocessing for classification in noisy healthcare data," *Machine Learning for Healthcare Technologies*, vol. 2, no. 33, p. 2016, 2016.

[66] P. Cao, X. Li, K. Mao, F. Lu, G. Ning, L. Fang, and Q. Pan, "A novel data augmentation method to enhance deep neural networks for detection of atrial fibrillation," *Biomedical Signal Processing and Control*, vol. 56, p. 101675, 2020.

[67] T. Li, M. Bolic, and P. M. Djuric, "Resampling methods for particle filtering: classification, implementation, and strategies," *IEEE Signal processing magazine*, vol. 32, no. 3, pp. 70–86, 2015.

[68] Y. Liu, Q. Li, K. Wang, J. Liu, R. He, Y. Yuan, and H. Zhang, "Automatic multi-label ecg classification with category imbalance and cost-sensitive thresholding," *Biosensors*, vol. 11, no. 11, p. 453, 2021.

[69] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[70] X. Liu, H. Wang, Z. Li, and L. Qin, "Deep learning in ecg diagnosis: A review," *Knowledge-Based Systems*, vol. 227, p. 107187, 2021.

[71] F. Murat, F. Sadak, O. Yildirim, M. Talo, E. Murat, M. Karabatak, Y. Demir, R.-S. Tan, and U. R. Acharya, "Review of deep learning-based atrial fibrillation detection studies," *International journal of environmental research and public health*, vol. 18, no. 21, p. 11302, 2021.

[72] S. Ogbomo-Harmitt, M. Muffoletto, A. Zeidan, A. Qureshi, A. P. King, and O. Aslanidi, "Exploring interpretability in deep learning prediction of successful ablation therapy for atrial fibrillation," *Frontiers in Physiology*, vol. 14, p. 1054401, 2023.

Food Data in the Semantic Web: A Review of Nutritional Resources, Knowledge Graphs, and Emerging Applications

1st Darko Sasanski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
darko.sasanski@finki.ukim.mk

2nd Riste Stojanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
riste.stojanov@finki.ukim.mk

***Abstract*—This comprehensive review explores food data in the Semantic Web, highlighting key nutritional resources, knowledge graphs, and emerging applications in the food domain. It examines prominent food data resources such as USDA, FoodOn, FooDB, and Recipe1M+, emphasizing their contributions to nutritional data representation. Special focus is given to food entity linking and recognition techniques, which enable integration of heterogeneous food data sources into cohesive semantic resources. The review further discusses food knowledge graphs, their role in semantic interoperability, data enrichment, and knowledge extraction, and their applications in personalized nutrition, ingredient substitution, food-drug and food-disease interactions, and interdisciplinary research. By synthesizing current advancements and identifying challenges, this work provides insights to guide future developments in leveraging semantic technologies for the food domain.**

***Keywords*—Nutritional Profiles, Semantic Resources, Knowledge Graphs, Food Entity Linking, Personalized Nutrition, Ingredient Substitution, Semantic Web**

I. Introduction

The vast and complex landscape of food-related data presents both challenges and opportunities in the domains of personalized nutrition, health research, and food safety. With the rise of semantic web technologies, food knowledge graphs (KGs) have emerged as powerful tools to organize, integrate, and make sense of diverse food data sources. These graphs enable the effective linking of heterogeneous datasets, including nutritional databases, recipes, food ontologies, and food science research, providing a comprehensive framework for understanding the relationships between food entities.

Key resources such as the USDA National Nutrient Database [1], FoodOn ontology [2], FooDB [3], and Recipe1M+ [4] have laid the foundation for the development of food knowledge graphs. These resources provide structured and standardized representations of food items, ingredients, and their nutritional properties, which can be leveraged to address a wide range of challenges in the food domain. The integration of these diverse datasets is crucial for applications such as personalized nutrition, ingredient substitution, food-disease interactions, and food safety.

Central to the development of food knowledge graphs is the use of food entity recognition (NER) and linking (NEL) techniques. These methods enable identification and connection of food-related data across multiple datasets, ensuring consistency and enhancing data interoperability. Such techniques are essential for creating more accurate and reliable systems that can inform personalized nutrition strategies, improve ingredient substitution recommendations, and support interdisciplinary research in the food-health domain.

This paper is structured as follows. The first section examines key food data resources, including USDA, FoodOn, FooDB, and Recipe1M+, providing a foundation for understanding the food data landscape. The second section explores food entity linking and recognition techniques, which enable the integration of heterogeneous data sources into unified semantic structures. The third section focuses on food semantic resources and knowledge graphs, emphasizing their role in enhancing data interoperability and knowledge extraction. The fourth section discusses various applications, such as personalized nutrition, ingredient substitution, and food-drug interactions, demonstrating the impact of these semantic resources. Finally, the conclusion synthesizes key insights, highlights current challenges, and outlines future directions for advancing semantic technologies in the food domain.

II. Data Resources in the Food Domain

This section provides an overview of the primary data resources in the food domain, highlighting their key characteristics, scope, and relevance to food science, nutrition, and related research areas. These resources serve as essential building blocks for developing semantic representations, enabling comprehensive analysis and integration of food-related data across various disciplines.

Recipe1M+ is a large-scale dataset containing over one million cooking recipes paired with approximately 13 million food images, introduced in [4]. Each recipe includes structured data such as titles, ingredients (with quantities and units), step-by-step instructions, and metadata like cooking times and

nutritional information. This rich dataset has become a foundational resource in computational gastronomy, enabling cross-modal learning, image-recipe retrieval, and various NLP tasks such as named entity recognition and ingredient substitution, fostering advancements in food informatics and semantic data integration.

USDA National Nutrient Database is a comprehensive database developed by the United States Department of Agriculture, consolidating several food composition resources like the Food and Nutrient Database for Dietary Studies (FNDDS) and the USDA Branded Food Products Database. Launched in 2019, it provides detailed nutrient profiles for a wide range of foods, from raw ingredients to branded products. With categories like Foundation Foods, Experimental Foods, and Standard Reference, FoodData Central supports applications in dietary assessment, public health, and nutritional research, making it an essential resource for professionals and researchers in food and nutrition [1].

In addition to the USDA's FoodData Central, numerous countries have developed their own Food Composition Databases (FCDBs) to support national dietary assessments and public health initiatives. For instance, the European Food Information Resource (EuroFIR) provides access to food composition data from multiple European countries, enabling cross-border nutritional research and policy development [5], [6].

FoodOn is an open-source ontology developed to standardize food-related data across various domains, including agriculture, nutrition, and food safety. It provides a controlled vocabulary for naming parts of animals, plants, fungi, and derived food products, enabling data integration and interoperability. By reusing terms from other ontologies, such as environmental terms from ENVO and agricultural terms from AGRO, FoodOn supports FAIR data annotation and sharing objectives across diverse sectors. This harmonization enhances global food traceability, quality control, and data integration, making it a vital resource for research and applications in food informatics [2].

FooDB is an open-access database providing comprehensive information on the chemical constituents of foods, including both macronutrients and micronutrients. It contains data on over 28,000 chemicals found in approximately 1,000 raw or unprocessed food products. Each chemical entry includes detailed compositional, biochemical, and physiological information, such as nomenclature, structure, chemical class, physicochemical properties, food sources, color, aroma, taste, physiological effects, and concentrations in foods. This resource is invaluable for researchers, dietitians, food scientists, and educators, offering a wealth of information on the chemical makeup of foods and their implications for human health [3].

SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) is a comprehensive clinical terminology system that includes codes for various health-related concepts, including food-related conditions. For example, it encompasses codes for food allergies and intolerances and various types of foods [7].

The Hansard corpus refers to the official verbatim report of debates in the UK Parliament, which often includes discussions on food-related policies and issues. Researchers have utilized the Hansard taxonomy to annotate food entities in various corpora, such as the FoodBase corpus, which contains annotated food entities extracted from recipes [8], [9].

FooDis is a crucial resource developed for extracting food-disease interactions from biomedical literature using advanced natural language processing techniques. By identifying and validating relationships between foods and health conditions, FooDis enables researchers to uncover potential cause-and-effect links that were previously difficult to detect. Its high precision in matching known associations makes it an essential tool for advancing our understanding of how diet influences disease and for enabling nutritional research and healthcare applications [10].

DrugBank is a comprehensive database that contains detailed information on over 15,000 drugs, including their chemical properties, mechanisms of action, and therapeutic uses. Although its primary focus is on pharmacological data, DrugBank also includes information on food-drug interactions, highlighting how certain foods can influence drug metabolism, absorption, and efficacy. This data is particularly valuable for research in nutritional genomics and personalized medicine, where understanding the interplay between food and drug interactions can improve treatment outcomes and patient safety [11].

AllergenOnline is a peer-reviewed, curated database designed to identify proteins that may present a potential risk of allergenic cross-reactivity. It serves as a valuable tool for assessing the safety of proteins introduced into foods through genetic engineering or processing methods [12].

The Cooking Ontology is a structured framework designed to represent key concepts in the culinary domain, including actions, food items, recipes, and utensils. It serves as a foundational resource for enhancing natural language processing capabilities in culinary applications [13].

The datasets and resources discussed throughout this section, summarized in Table I, provide essential data for advancing research in the food domain. These resources play a critical role in various applications, such as personalized nutrition and food composition analysis. To utilize these datasets effectively, it is vital to perform accurate entity linking, ensuring the correct identification and alignment of food-related entities across multiple resources. The next section delves into the methodology of entity linking and recognition in the food domain.

III. Food Entity Linking and Recognition

Entity Linking (EL) and Entity Resolution (ER) are essential tasks in Natural Language Processing (NLP) that involve identifying, disambiguating, and mapping entities in text to a standardized set of identifiers or canonical forms. EL connects mentions of entities to their corresponding entries in a knowledge base, while ER merges different mentions referring

TABLE I: Summary of Key Food Datasets and Resources

Dataset/Resource	Key Features	Primary Application
Recipe1M+	1M+ recipes, 13M images, structured data	Cross-modal learning, ingredient substitution
USDA National Nutrient Database	Nutrient profiles for foods, food categories	Dietary assessment, public health, nutritional research
FoodOn	Controlled vocabulary, FAIR data sharing	Food traceability, quality control, data integration
FooDB	Chemical consituents of foods, 28,000+ chemicals, biochemical data	Food composition research
SNOMED CT	Health-related food terms, allergy data	Health research, food allergies
The Hansard corpus	UK Parliament debates, food policies	Food entity annotation, policy research
FooDis	Food-disease interactions	Nutritional research, healthcare
DrugBank	15,000+ drugs, food-drug interactions	Nutritional genomics, personalized medicine
AllergenOnline	Peer-reviewed allergenic protein database	Allergen research, food safety
The Cooking Ontology	Structured framework for culinary domain	Culinary applications, NLP enhancement

to the same real-world entity. Both processes enhance data integration, information extraction, and semantic understanding of textual data [14], [15].

Over the past decade, food Named Entity Recognition (NER) and Named Entity Linking (NEL) have evolved significantly, transitioning from rule-based methods to advanced machine learning techniques.

Early systems focused on rule-based methods to extract food-related information from various sources:

- **drNER**: A rule-based method designed to extract dietary information from evidence-based recommendations, enabling structured analysis of nutritional guidelines [16].
- **FoodIE**: A rule-based method for extracting food entities from unstructured data, utilizing computational linguistics and semantic information [17].
- **StandFood**: A classification method emphasizing the lexical similarity of food entities, linking them to the FoodEx2 database by the European Food Safety Authority (EFSA) [18].

The introduction of the FoodBase annotated corpus [9] marked a key advancement, enabling the use of machine learning techniques:

- **BuTTER**: A machine learning-based method using a Bidirectional Long Short-Term Memory (BiLSTM) network combined with Conditional Random Fields (CRF) for identifying food entities [19].
- **FoodNER**: A set of models fine-tuned from pretrained BERT architectures to extract and annotate food entities in diverse contexts, representing a major step forward in food information extraction [20].

To assist experts in navigating different food standards and improving interoperability, **FoodViz** was developed as a web-based framework for presenting food annotation results from NLP and machine learning pipelines [21].

The study "Zero-shot evaluation of ChatGPT for food named-entity recognition and linking" evaluates the performance of ChatGPT-3.5 and ChatGPT-4 in identifying and linking food-related entities without prior task-specific training. Results show both models perform better with consumption data than scientific literature, with ChatGPT-4 showing slight improvement over ChatGPT-3.5. However, both models struggle with linking entities to standardized identifiers, indicating the need for further refinement [22].

Despite significant advancements in food NER and NEL, challenges such as variability in food terminology, the scarcity of comprehensive annotated datasets, and the complexity of food compositions remain. Addressing these issues is essential for improving the accuracy and reliability of food information extraction systems [22], [23].

Beyond entity recognition and linking, integrating data from heterogeneous food sources introduces additional layers of complexity. For instance, aligning resources like USDA and FooDB is non-trivial: while both offer structured nutritional information, they differ significantly in terminology, unit representation (e.g., mg vs. g), ingredient granularity (e.g., milk vs. whole cow's milk), and conceptual hierarchies. Ontologies such as FoodOn and SNOMED CT further complicate alignment due to overlapping, yet semantically nuanced, definitions. These differences highlight the need for sophisticated linking approaches that go beyond lexical similarity, underscoring the importance of harmonization frameworks and robust evaluation techniques.

IV. Food Semantic Resources and Knowledge Graphs

Food semantic resources and knowledge graphs integrate diverse food-related data, including nutritional, molecular, and culinary information. By linking datasets from food ontologies, nutritional databases, and recipes, they enable comprehensive analysis and interdisciplinary research in personalized nutrition, food safety, and sustainability. This section overviews key semantic resources and knowledge graphs in the food domain, highlighting their contributions and impact on food science and technology.

FoodOntoMap is a resource developed to link food concepts across various food ontologies, enabling normalization and integration of food-related data. By mapping food concepts extracted from recipes to semantic tags from multiple food ontologies, it enables a unified representation of food information across different domains [24].

The dataset includes food concepts extracted from 22,000 recipes sourced from AllRecipes [25]. Each ingredient is annotated using ontologies available in BioPortal [26], such as FoodOn, OntoFood, SNOMED CT, and the Hansard Corpus.

This mapping links distinct food ontologies, enabling applications that analyze relationships between food systems, human health, and the environment.

FoodOntoMap addresses the challenge of integrating diverse food ontologies, each developed for specific applications, by providing a unified framework. This interoperability supports the development of tools for studying food-health-environment interactions and enhances the consistency of food data representation.

The resource has been further expanded with additional ontologies, broadening its coverage and reinforcing its role in semantic food data integration. By improving interoperability, FoodOntoMap advances research and application development in nutrition science, food safety, and sustainability.

In summary, FoodOntoMap serves as a crucial tool for linking and normalizing food concepts across ontologies, enabling a more comprehensive understanding of food-related data.

FoodKG is a comprehensive knowledge graph designed to integrate diverse food-related data into a unified structure, connecting food concepts and their relationships across various domains. It provides a semantic framework for understanding food systems, human health, and the environment by linking food entities to a wide range of data, including nutritional information, recipes, and ontologies [27].

The construction of FoodKG integrates multiple data sources, such as recipe data from Recipe1M+, nutrient information from the USDA, and semantic resources like FoodOn, forming a cohesive knowledge graph. This integration addresses challenges such as standardizing food data, merging fragmented sources, and enhancing data accessibility. As a result, FoodKG supports applications in personalized nutrition, recipe recommendation, and food-disease research.

FoodKG has enabled the development of practical tools, including a SPARQL-based service that recommends recipes based on available ingredients and constraints like allergies, as well as a cognitive agent for natural language question answering on the knowledge graph.

In summary, FoodKG is an essential resource for connecting and integrating food-related data, improving interoperability, and enabling applications that enhance dietary choices, health outcomes, and environmental sustainability.

TABLE II: Inconsistent Mappings for Cheese Ingredients [15]

Ingredient	USDA Class
1% fat cottage cheese	Blue Cheese
2% cheddar cheese	Blue Cheese
Anejo cheese	Blue Cheese
Better Than Cream Cheese cream cheese substitute	Blue Cheese
Cotija cheese	Blue Cheese
Greek feta cheese	Blue Cheese

In recent advancements, the preprocessing steps for entity resolution (ER) outlined in the FoodKG framework have undergone significant refinement [15]. Initially, the mappings within the knowledge graph were found to be unreliable, which hindered the accuracy and effectiveness of any subsequent research or analysis, as evidenced in Table II. This issue highlighted the need for a more robust and accurate alignment process. As part of the effort to address these limitations, this study [15] has not only reworked the alignment methodology but also significantly expanded the knowledge graph to incorporate critical information on the chemical compositions and health effects of food constituents by integrating the FooDB dataset. This addition enables a richer understanding of food data, linking nutritional and chemical information with health outcomes, an essential aspect for advancing both food science and health research. The comparison of linkage statistics between the original FoodKG and the newly expanded knowledge graph is presented in Table III, demonstrating substantial improvements in coverage and data integration.

TABLE III: Linkage Statistics [15]

Statistics	FoodKG Count	New KG Count
Dataset Linkages		
Number of links to USDA	5934	10417
Number of links to FoodOn	2192	10586
Number of links to FooDB	0	9377
Additional Linkage Stats		
Percentage of ingredients linked to USDA	32.51%	90.87%
Percentage of ingredients linked to FoodOn	12.01%	92.35%
Percentage of ingredients linked to FooDB	0%	81.80%

TABLE IV: Comparison of the mappings [15]

	Our Mapping	FoodKG's Mapping
True Positives	1985	1078
False Positives	618	1525
Precision (%)	76.25	41.41

To ensure the reliability of the newly generated mappings, a rigorous evaluation process was implemented. Students annotated a selected subset of these mappings, who provided detailed feedback on the quality and relevance of the links. As shown in Table IV, the feedback strongly favored the newly generated mappings, indicating a significant improvement in both accuracy and utility.

The integration of multiple aspects of the food domain, including nutritional information, chemical compositions, and health effects, into a unified knowledge graph represents a significant milestone. For the first time at this scale, these diverse aspects of food data are encoded into a single resource, enabling comprehensive interdisciplinary research. This work lays the foundation for advanced applications in personalized nutrition, disease prevention, and sustainable food systems, providing researchers with a robust tool for a wide range of studies.

V. Applications of Food Semantic Resources

Food semantic resources, such as knowledge graphs and linked datasets, have become essential tools in addressing complex challenges in food science, nutrition, and related fields. By integrating diverse data sources, these resources enable advanced applications ranging from personalized nutrition and dietary analysis to food safety, sustainability, and

health research. This section explores key applications that leverage food semantic resources to enhance research, improve decision-making, and foster interdisciplinary collaboration.

A basic use case for food semantic resources is their application in personalized recipe recommendation systems. By leveraging SPARQL queries, these systems can filter recipes based on dietary preferences and constraints. For example, users can search for recipes that include or exclude specific ingredients, such as vegetables, meat, or allergens. Nutritional filtering can also be applied, enabling users to find recipes low in sugar, fat, or calories, which is particularly useful for managing health conditions like diabetes or hypertension.

These use cases are demonstrated in the FoodKG paper, where SPARQL queries were employed to filter recipes by ingredients and nutritional values, addressing health-related dietary restrictions. In addition to ingredient and nutritional filtering, these systems can support dietary preferences by categorizing ingredients based on food types, such as vegan, vegetarian, or halal, ensuring recipes align with users' lifestyle choices. The integration of these use cases within food knowledge graphs enables personalized, health-conscious recipe recommendations, helping users make informed dietary choices.

Two approaches highlight the potential of using natural language for querying semantic resources. The first, from the FoodKG paper, demonstrates Knowledge Base Question Answering (KBQA) applied to food-related knowledge graphs. Users can pose natural language questions, and the system retrieves answers from the graph, simplifying interaction with complex data. The paper also introduced a synthetic Q&A dataset to test this, with questions ranging from simple ones like "How much sugar is in cheese, cream, fat free?" to more complex constraints-based queries like 'What Laotian dishes can I make with sugar, water, oranges?' This method enhances the accessibility of food knowledge graphs, enabling intuitive querying based on natural language.

The second approach, in Bridging the Gap: Generating a Comprehensive Biomedical Knowledge Graph Question Answering Dataset, presents a similar method applied to the biomedical domain. They created a Q&A dataset based on the PrimeKG graph, offering over 83,000 question-answer pairs that could be easily reproduced in the food domain. Both methods showcase how natural language can make complex knowledge graphs more accessible, enabling intuitive and efficient querying across different domains and highlighting the significant benefit of simplifying the exploration of large, complex datasets [28].

A key application of food knowledge graphs is ingredient substitution, which is essential for meeting dietary or health needs. The paper Identifying Ingredient Substitutions Using a Knowledge Graph of Food demonstrates how food knowledge graphs, combined with semantic data and word embeddings, can identify suitable ingredient substitutes based on factors like nutritional content and food classifications [29].

Food knowledge graphs enable automatic generation of ingredient alternatives to accommodate dietary restrictions such as allergies, low sugar diets, or vegan preferences. The system ranks substitutions based on their compatibility with the original recipe, offering flexibility for users to adapt meals to their health goals or ingredient availability. This approach allows individuals to modify familiar meals rather than following rigid meal plans, promoting personalized nutrition.

Despite the potential, current knowledge graphs, like FoodKG, may overlook certain aspects of ingredients, such as their role in recipes, taste, or texture, which can impact substitution accuracy. However, the newly generated semantic resource from [15], which includes information about the molecular composition of ingredients, could enhance substitution accuracy. Since ingredients with similar molecular structures are likely to share comparable taste and texture, as discussed in [30], this resource has the potential to suggest substitutions that preserve the original recipe's flavor and structure. This approach would improve the accuracy of ingredient recommendations, ensuring that dietary restrictions are met while maintaining the sensory qualities of the original dish.

Integrating the FoodOn ontology into food knowledge graphs enhances various applications by categorizing food entities based on their origin, processing, and supply chain relationships [2]. This improves food traceability, enabling identification of contamination sources and affected products, recipes, or distribution points, thereby reducing economic losses and health risks.

Integrating FoodOn with species and anatomical ontologies enables identification of toxic compounds and analysis of interactions between food components and medications. By linking food chemicals to the Comparative Toxicogenomics Database (CTD), researchers can assess how dietary compounds influence gene behavior and contribute to disease risk. This approach not only strengthens food safety measures but also helps tailor dietary recommendations to individual health needs [2], [31], [32].

Additionally, aligning with resources like FooDB and FoodOn helps establish food-disease relationships, enabling personalized recipe recommendations tailored to specific health needs [10]. In summary, integrating FoodOn strengthens food traceability, safety, and personalized nutrition through a standardized framework for analyzing food entities and their properties.

VI. Conclusion

In this review, we explored the evolving landscape of food data and the pivotal role of semantic technologies, particularly food knowledge graphs—in transforming how food-related information is organized, integrated, and applied. We examined key resources such as USDA, FoodOn, FooDB, and Recipe1M+, highlighting their foundational roles in food knowledge representation. We also reviewed food entity recognition and linking techniques, emphasizing how they enhance interoperability and enable applications in personalized nutrition, ingredient substitution, and food safety.

While the integration of these resources is already driving innovations in personalized nutrition, improving health outcomes, and enabling interdisciplinary research in the food-health domain. However, challenges remain, including the need for more comprehensive datasets, standardization, and effective methods for handling data inconsistencies.

Unlike prior reviews that focus on isolated datasets or applications, this work offers a comprehensive synthesis across nutritional databases, ontologies, and knowledge graphs. It is, to our knowledge, the first to jointly analyze food entity recognition techniques, dataset alignment strategies, and semantic applications in food science. This integrated perspective also highlights the emerging role of large language models (LLMs) in supporting semantic alignment and content enrichment.

Looking ahead, the continued development of food knowledge graphs and the refinement of entity recognition and linking techniques will further enhance the accuracy and utility of these resources. Future research should focus on expanding the scope of food knowledge graphs, incorporating more diverse food data, and improving the scalability of these systems for broader adoption. By doing so, we can unlock even greater potential for personalized health interventions, enhanced food safety, and more sustainable food systems.

References

[1] D. Haytowitz, L. Lemar, P. Pehrsson, J. Exler, K. Patterson, R. Thomas, M. Nickle, J. Williams, B. Showell, and M. Khan, "Usda national nutrient database for standard reference, release 24," *US Department of Agriculture: Washington, DC, USA*, 2011.

[2] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. Brinkman, and W. W. Hsiao, "Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration," *npj Science of Food*, vol. 2, no. 1, p. 23, 2018.

[3] D. Wishart, "Foodb: the food database," 2018.

[4] J. Marın, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba, "Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 187–203, 2021.

[5] H. Greenfield and D. A. Southgate, *Food composition data: production, management, and use*. Food & Agriculture Org., 2003.

[6] S. Westenbrink, M. Oseredczuk, I. Castanheira, and M. Roe, "Food composition databases: the eurofir approach to develop tools to assure the quality of the data compilation process," *Food chemistry*, vol. 113, no. 3, pp. 759–767, 2009.

[7] K. Donnelly *et al.*, "Snomed-ct: The advanced terminology and coding system for ehealth," *Studies in health technology and informatics*, vol. 121, p. 279, 2006.

[8] Hansard. Hansard corpus. https://www.hansard-corpus.org. Accessed: February 2025.

[9] G. Popovski, B. K. Seljak, and T. Eftimov, "Foodbase corpus: a new resource of annotated food entities," *Database*, vol. 2019, p. baz121, 2019.

[10] G. Cenikj, T. Eftimov, and B. K. Seljak, "Foodis: A food-disease relation mining pipeline," *Artificial intelligence in medicine*, vol. 142, p. 102586, 2023.

[11] C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. Chin, S. A. Strawbridge *et al.*, "Drugbank 6.0: the drugbank knowledgebase for 2024," *Nucleic acids research*, vol. 52, no. D1, pp. D1265–D1275, 2024.

[12] R. E. Goodman, M. Ebisawa, F. Ferreira, H. A. Sampson, R. van Ree, S. Vieths, J. L. Baumert, B. Bohle, S. Lalithambika, J. Wise *et al.*, "Allergenonline: a peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity," *Molecular nutrition & food research*, vol. 60, no. 5, pp. 1183–1198, 2016.

[13] F. Batista, J. P. Pardal, P. V. N. Mamede, and R. Ribeiro, "Ontology construction: cooking domain," *Artificial Intelligence: Methodology, Systems, and Applications*, vol. 41, no. 1, p. 30, 2006.

[14] IBM Docs. Entity resolution. https://www.ibm.com/docs/en/iii/9.0.0?topic=insight-entity-resolution. Accessed: 5-May-2024.

[15] D. Sasanski, A. Todorovski[1], B. Trpeski, D. Trajanov, T. Eftimov, and R. Stojanov, "Aligning food ingredients with multiple," in *ICT Innovations 2024. TechConvergence: AI, Business, and Startup Synergy: 16th International Conference, ICT Innovations 2024, Ohrid, North Macedonia, September 28–30, 2024, Proceedings*, vol. 2436. Springer Nature, 2025, p. 19.

[16] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PloS one*, vol. 12, no. 6, p. e0179488, 2017.

[17] G. Popovski, S. Kochev, B. Korousic-Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction." *ICPRAM*, vol. 12, p. 915, 2019.

[18] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Standfood: standardization of foods using a semi-automatic system for classifying and describing foods according to foodex2," *Nutrients*, vol. 9, no. 6, p. 542, 2017.

[19] G. Cenikj, G. Popovski, R. Stojanov, B. K. Seljak, and T. Eftimov, "Butter: Bidirectional lstm for food named-entity recognition," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 3550–3556.

[20] R. Stojanov, G. Popovski, G. Cenikj, B. Koroušić Seljak, and T. Eftimov, "A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation," *Journal of Medical Internet Research*, vol. 23, no. 8, p. e28229, 2021.

[21] R. Stojanov, G. Popovski, N. Jofce, D. Trajanov, B. K. Seljak, and T. Eftimov, "Foodviz: Visualization of food entities linked across different standards," in *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part II 6*. Springer, 2020, pp. 28–38.

[22] M. Ogrinc, B. Koroušić Seljak, and T. Eftimov, "Zero-shot evaluation of chatgpt for food named-entity recognition and linking," *Frontiers in nutrition*, vol. 11, p. 1429259, 2024.

[23] U. Akujuobi, S. Liu, and T. R. Besold, "Revisiting named entity recognition in food computing: enhancing performance and robustness," *Artificial Intelligence Review*, vol. 57, no. 9, p. 241, 2024.

[24] G. Popovski, B. Korousic-Seljak, and T. Eftimov, "Foodontomap: Linking food concepts across different food ontologies." in *KEOD*, 2019, pp. 195–202.

[25] S. Groves, "How allrecipes. com became the worlds largest food/recipe site. roi of social media (blog)," Tech. Rep, Tech. Rep., 2013.

[26] 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.

[27] S. Haussmann, O. Seneviratne, Y. Chen, Y. Ne'eman, J. Codella, C.-H. Chen, D. L. McGuinness, and M. J. Zaki, "Foodkg: a semantics-driven knowledge graph for food recommendation," in *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*. Springer, 2019, pp. 146–162.

[28] X. Yan, P. Westphal, J. Seliger, and R. Usbeck, "Bridging the gap: Generating a comprehensive biomedical knowledge graph question answering dataset," in *ECAI 2024*. IOS Press, 2024, pp. 1198–1205.

[29] S. S. Shirai, O. Seneviratne, M. E. Gordon, C.-H. Chen, and D. L. McGuinness, "Identifying ingredient substitutions using a knowledge graph of food," *Frontiers in Artificial Intelligence*, vol. 3, p. 621766, 2021.

[30] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, "Flavor network and the principles of food pairing," *Scientific reports*, vol. 1, no. 1, p. 196, 2011.

[31] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics database: update 2017," *Nucleic acids research*, vol. 45, no. D1, pp. D972–D978, 2017.

[32] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D668–D672, 2006.

Robotics, Electrical Engineering, and Tourism in Macedonia

Vangel Bozinoski
Section of Geography
University of Tourism and Management
Skopje, Macedonia
v.bozinoski@utms.edu.mk

Stevo Bozinovski
Department of Computer Science and Mathematics
South Carolina State University
Orangeburg, SC, USA
sbozinovski@scsu.edu

***Abstract*—The paper discusses possibility of tourist offering related to robotics and electrical engineering in Macedonia. We are especially interested in events and people related to those topics, which might contribute to a narrative for the tourist offering.**

Keywords—Robotics, electrical engineering, relevant places, Mihajlo Pupin, Nikola Tesla, tourist offering

I. Introduction

A part of tourist offering of a country are the stories that are connected to events and individuals. In addition to events in history and culture, countries often relate themselves with technical events. For example, in North Carolina in the USA, the vehicles have license plates where it is written "First in Flight", relating North Carolina to the first flight of the Wright Brothers.

Here we are interested in a narrative about robotics and electrical engineering with related events and individuals in Macedonia. We believe that there is a lack of such tourist offerings. The result of our research are three places which are connected with robotics and electrical engineering, and could offer interesting insights to tourists. Here we will restrict ourselves to Skopje, Vevcani near Struga, and Izvor near Veles.

II. Skopje

Arguably the most important event in Robotics and Electrical Engineering in Skopje is the international recognition named IEEE Milestone, for the important pioneering scientific event, namely the first control of a robot using brain signals.

The international organization, Institute of Electrical and Electronic Engineers (IEEE) placed a monument in Macedonia for the mentioned pioneering event. The monument is in the form of a bronze plaque placed on a wall. The monument was unveiled by the IEEE President, Dr Saifur Rahman, and the Macedonian president Dr. Stevo Pendarovski, on October 10, 2023. The monument was placed, and is still standing, on the wall of the annex building of the Faculty of Electrical Engineering and Information Technologies (FEIT). Figure 1 shows the plaque on the wall of the FEIT building in the moment after unveiling the plaque.



Figure 1. The monument on the first control of a robot using brain signals placed in the building of the FEIT. Below the monument from left to right: the representative of the IEEE Macedonia section Goga Cvetkovski, Mihail Sestakov, IEEE President Saifur Rahman, Macedonan president Stevo Pendarovski, Stevo Bozinovski, and Liljana Bozinovska.

The pioneering event, first control, of a robot using EEG signals, happened in 1988 in the Laboratory of Intelligent Machines and Bioinformation Systems of the Electrical Engineering Faculty (ETF), which was the antecedent to FEIT. Three people from Macedonia created this contribution to science: electrical engineers Stevo Bozinovski and Mihail Sestakov, and medical doctor Liljana Bozinovska. They published a paper about it in 1988 [1].

This event could be one of the tourist offerings in a tour through Skopje. A tourist group might be lead to the IEEE plaque, telling stories about how it happened that the first in the world control of a robot using brain signals took place indeed in Macedonia.

It should be noted that the nearby countries in the region have no IEEE Milestone achievements for a pioneering event in science. There are recognition for contribution in a *regional development*, such as hydro power plants in Croatia (Sibenik 1895, Austro-Hungary) and Serbia (Vucje in 1903), while Hungary has a recognition for the Budapest metro. In the region (between and including Austria, Greece, and Romania) only Macedonia has an IEEE Milestone recognition for a *pioneering scientific achievement.*

Another robotics achievement that could be narrated to tourists is the first control of a robot using speech commands. In 1986 the first in the world robot was controlled by five Macedonian language commands: trgni, levo, desno, nazad and stoj [2][3]. To tourists should know that the "first robot in the world which understood speech commands was a Macedonian".

Yet another robotics achievement is the first control of a robot using multitasking programming. In multitasking programming a programmer divides the programs into several subprograms (tasks) and manages communication between them, such that the tasks contribute toward the fulfillment of the purpose of the program. This was achieved in 1983 by the ETF, where its students were the first in the world to write and execute a multitasking program. This was enabled by the computer IBM Series/1, which the students used [4][5].

From another tourist perspective toward Skopje, let us first mention that it is believed that all Macedonian settlements are established in the Neolithic era. Skopje is a good example of a settlement dating from the Neolithic era. According to information given by Maria Gimbutas [6], Macedonian settlements are the oldest in Europe, dating back to earlier than 8750 BC. It is also interesting that the names of the settlements, mountains, valleys, have not been changed since the time of the Neolithic revolution (migratory society to sedentary, agricultural society) [7].

Maybe the most attractive tourist offering in Skopje is the centre of the city, with the architectural composition between the monuments of Alexander of Macedon and his father Philip II of Macedon, with the river Vardar and Justinian bridge between them. Here under "city center" is considered the center of the Justinian bridge. Many monuments make up the composition, and it can actually be viewed as an open-space museum. [8]. The composition offers various sculptures, one of them being the first European king Karan, who put his profile on minted coins. The small clay figurine known as "Adam from Macedonia" is proclaimed as the most important artifact for Europe for the year 2000.

Another historically important offering is the Church of the Holy Savior (orig. Свети Спас) where lie the remains of the Macedonian revolutionary and philosopher Goce Delcev.

Let us also mention Mother Teresa's house, which is a part of the architectural opus by the first author of this text. It is currently the most visited museum in Macedonia.

Let us also mention that in the Second World War (WW2) the city of Skopje was liberated by Macedonian part of Yugoslavian army. After a fight against Germans a flag was placed on the Officers' Home in Skopje, by Penka Arsovska (Kralevska), on November 13, 1943. Before that Bulgaria which was on the Nazi side, and ruled Macedonia in the WW2, was defeated by Red Army.

III. VEVCANI

In tourist offering, Vevcani is a well-known place in Macedonia [9][10]. It is known for its carnival which takes place in January each year. It is also known for its wild natural water springs.

In 1987 there was an intention of the government to take the springs from Vevcani and distribute it to other settlements, so that Vevcani would have no possession over its springs. There was an uprising from the villagers against such an intention. The police was engaged with electric sticks, which were applied for the first time in Macedonia, on Vevcani's women. The government applied imprisonment, disqualifications, and other types of pressure. Vevcani fought for three years and in the end the government apologized for this type of pressure. Vevcani defended its springs from converting them into a water reservoir, and preserved the natural beauty of the springs [11].

Vevcani names itself the Republic of Vevcani, and has its own passport, and its own money (the Licnik), This started on January 12, 1992, when Vevcani organized a referendum to protect the unitarity of Macedonia, due to various disintegration-oriented political movements at that time [11]. Today it is a part of the tourist offering by Vevcani.

Vevcani and people from Vevcani, the Vevcans, are proud of their culture. An example is a book written by Jordanka Popeska [12]. The book is an effort toward a grammar of the Vevcani speech.

Importantly, in a church that was recently unearthed, a Masonic symbol was found, suggesting that a branch of the Masonic movement has had an origin in area of Vevcani. The mountain area around Vevcani is also famous for its Cave churches.

Vevcani was engaged in the Ilinden Uprising in 1902-03 against the Ottoman Empire for independent Macedonia. A participant was Kota Kostojcinova Pupina [13]. We will mention that the mother of the authors of this paper is from Kostojcinoski family from Vevcani.

For electrical engineering it is important that Vevcani is ancestral place of Mihajlo Pupin (1854-1935). He was a professor at Columbia University in the USA, and is known for his numerous patents, including an innovation of so called Pupin Coils, which greatly extending the range of long distance telephone communication.

Mihajlo Pupin's ancestors settled in Vevcani in 1690. After about 150 years, a member of the family (Konstantin) moved to Military Border of the Habsburg Empire. At that time a Military Border was created by Habsburg Empire toward the Ottoman Empire. Many people from Ottoman empire were attracted to live in the Military Border, with an idea that they will protect the Habsburg Empire in the next invasions by the Ottoman Empire. Konstantin moved from Vevcani, Monastir Region of the Ottoman Empire, to Idvor, Military Border, where Mihajlo Pupin was born. There are still about 10 Pupin families living in Vevcani today. The neighborhood where they live now is called Pupini [11].

Figure 2 shows the house of the Pupin family in Vevcani. [15]



Figure 2. The Pupin family house in Vevcani

The dominant story today is that Pupin has a Serbian origin. His name was extended to Mihajlo Idvorski Pupin, emphasizing that he was born in Idvor. He also used his name Mihajlo Pupin, for example when received a Pulitzer Prize. Maybe a more explanatory name should be Mihajlo Vevcanski Idvorski Pupin, to better describe his family history.

In 2019 the foundation Selak from Serbia placed a monument of Mihajlo Pupin in Vevcani pointing out a relation of Mihajlo Pupin to Vevcani.

During his work in the USA, Mihajlo Pupin carried special love for Macedonia and Ohrid. He built the bell tower and donated a church bell to the church St. Clement Ohridski in Ohrid [13], which is still there. During the Paris Conference 1919 he urged for autonomy of Macedonia [16].

Like almost every settlement in Macedonia, Vevcani was founded in the early Neolithic era. There are several neolithic sites in and around Vevcani but only in the Plesec archaeological institutions one trench is open, rich with artefacts of a permanent settlement from the neolith until the end of the iron age. Vevcani is a settlement that lies on the famous Via Ignatia road, known locally as Via Sacra or King's Road. Historically, King's Road is a name for first international road built by the agreement between Macedonia and Persia in the 5th century BC. The remains of Via Ignatia were found above Vevcani.

Saint Paul had traversed this area and by a word of mouth left a gift for Macedonian early Christian believers. The gift was a 3 stones mark with the blood of Jesus from the Cross. After the official recognition of the Christian religion in the 4th century, big cathedrals were erected in this region.

IV. Izvor

Izvor is an old settlement. A Neolithic location of Izvor is named Dzuniver and is dated around 6000 years BC.

Izvor is currently not much present in the tourist offering. We would mention its water spring which is like a small lake. Figure 3 shows the Izvor's spring.



Figure 3. The water spring in Izvor

However, arguably, a well known electrical engineer and inventor Nikola Tesla (1856-1943) has his origins from this place, which makes Izvor a potential for a important tourist offering.

Tesla is known for many inventions. The most important is the rotating magnetic field and the use of the alternating current, which enabled the entire world to use electric energy. His work is the first hydro electric power plant on the Niagara river. For our story about robotics, it is important that he built and demonstrated the first remote control of a boat robot.

The dominant story today is that Tesla was born in the village of Smiljan, Lika, and has a Serbian origin, because his family was Orthodox Christian.

Another story is that at that time many families from Macedonia were moved to Lika, and part of Tesla's family moved from Izvor, Monastir Region, Ottoman Empire, to Smiljan, Military Border, Habsburg Empire.

A version of that story is that, according to Jaklina Blazevski, Tesla was not born in Smiljan, he was given from Macedonia, for adoption in Lika [17]. At that time Military Border welcomed for adoption children from Ottoman Empire. He was baptized in a church in Smiljan.

In our research we collected various stories about Tesla. There are two sets of sources in our research. One is from Internet and literature, the other is our personal communication with people.

On Internet, prof. Angelina Markus wrote: "Nikola Tesla is a Macedonian, born in the village of Izvor near Bogomila". With his Macedonian relatives he wrote letters in the Macedonian Language. The letters were found in the labs of the offices of Niagara. In one of the letters he wrote "I am equally proud of my Croatian and Macedonian Origin" [17].

In our research we visited Izvor and talked with the local people. A story came from a retired teacher, Mile Andreevski, who carries memories of his grandmother Stojna who was a teacher in old Yugoslavia. When Mile was learning about Nikola Tesla in school (around 1960), his grandmother told him that Tesla is indeed from Izvor, from their village [18].

Lidija Demnieva remembers that her grandfather Milan Demniev told her that Tesla is from Izvor [19].

V. Discussion

Basically there are two components of a tourist offering. The first contains the current natural beauties, current architecture, and culinary specifics of a place. The other important component is the historical narrative about places and events. In this paper we focused on some historical events related to robotics and electrical engineering that are connected to Macedonia.

We understand history as something connecting to a nation and often to politics. That is why different states and nations have different, often conflicting, histories. Because of conflicting histories, there might be conflicts between governments. As opposite to history, there is the History Science. The History Science is a single one, there are no two or more History Sciences. There might be various theories in the History Science, and usually one of the theories is the dominant theory. Others are revolutionary theories, or competitive theories. In any case, the History Science embraces all theories, expecting that in time some of them will prevail as dominant ones. Crucially, in the History Science there is no inquisition. There is not a case that one country or some organization will forbid another country or nation to tell its own history. The History Science is a science, while history itself might be a tool to control the thinking and behavior of another nation.

VI. Conclusion

Among other tourist offerings of Macedonia, there are stories related to robotics and electrical engineering that might be included in those offerings. In the future that material might be a part of a Macedonian museum for Science and Engineering

Here we researched the possibility of offering events that happened in the area of robotics and electrical engineering, emphasizing the places related to relevant people. We were interested in the Macedonian origins of Mihajlo Pupin and Nikola Tesla.

We hope that this will help the development of tourism in Macedonia, as well as of robotics and electrical engineering. It is important that younger generations be inspired by the contributions of the older ones.

References

[1] S. Bozinovski, M. Sestakov, L. Bozinovska,. “Using EEG alpha rhythm to control a mobile robot,” In G. Harris, C. Walker (eds.) Proc 10th Annual Conference of IEEE Engineering in Medicine and Biology Society, Track 17: Biorobotics, New Orleans, vol 3, pp. 1515-1516, 1988

[2] G. Grujovski, S. Bozinovski, "Realization of a system for recognition of isolated words and control of a mobile robot with speech commands (in Macedonian). Technical report, Project Adaptive Industrial Robots, 1986.

[3] G. Grujovski, S. Bozinovski, "Realization of a system for speech control of a mobile robot" (in Macedonian). In Proceedings of the 6th Yugoslav Symposium on Applied Robotics and Flexible Automation, Novi Sad, pp. 227–235, 1989

[4] S. Bozinovski, M. Sestakov, "Multitasking operating systems and their application in robot control" (In Macedonian) Proc. Workshop on Macedonian Informatics, Skopje, pp. 195-199, 1983

[5] S. Bozinovski, "Parallel programming for mobile robot control: Agent-based approach", IEEE Proc. Int. Symp. on Distributed Computing Systems, pp. 202-208, Poznan, 1994

[6] M. Gimbutas, The Civilization of the Goddess: The World of Old Europe. Harper, San Francisco, 1991

[7] V. Bozinoski Contribution Towards the History of Macedonian Architecture. (In Macedonian), Macedonian Matica, 2014

[8] V. Bozinoski. Off Balance Sheet (In Macedonan "Von Balansnata Evidencija", Published by the author, 2024

[9] C. Ristova, T. Angelkova-Petkova "Hotel development of cultural tourism elements: the case of Vevcani, Republic of Macedonia", Culture, Heritage and Tourism Development, Proc. Conf SITCON 2018, pp. 173-179

[10] N. Dimitrov "Tourist places, facilities and events that Macedonians should visit" International Journal of Economics, Management and Tourism Vol 4, No. 2, pp 28-42, 2024

[11] A. Popeski. Personal communication, 2025

[12] J. Popeska. On Vevcani in Vevcanian (original in Macedonian "Za Vevcani na Vevcanski"), Rojal Art, Skopje, 2021

[13] K. Aleksieska, Trite Univerziteti na prof.dr. Pavle Mitreski. Iris, 2018

[14] P. Mitreski. Mihajlo Pupin: From Vevcani to Idvor. Iris, 2020

[15] https://mymacedoniablog.com/sightseeing/ohrid-and-surroundings/vevchani/

[16] https://mk.wikipedia.org/w/index.php?title=Михајло_Пупин&oldid=5282937

[17] https://mactel.com.au/en/koj-e-golemiot-naucnik-nikola-tesla

[18] Mile Andreevski from Izvor, personal communication, 2024

[19] Lidija Demnieva, personal communication, 2025

Optimization of the Race Strategy in Formula 1

Marija Saveska
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University
Skopje, North Macedonia
marija.saveska@students.finki.ukim.mk

Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University
Skopje, North Macedonia
dimitar.trajanov@finki.ukim.mk

***Abstract*—Optimizing the race strategy in Formula 1 requires a comprehensive understanding of the factors that influence performance, including lap times, tyre choices, pit stops and weather conditions. This thesis analyzes race data from the 2024 season using Exploratory Data Analysis (EDA) to uncover patterns and relationships between key variables. By examining distributions, correlations and trends, EDA highlights how different factors interact and impact overall race outcomes. Machine learning techniques further refine this analysis by applying regression models to determine which variables have the greatest influence on lap time and position changes. The results show that data can be a key factor in making strategic decisions, giving a better representation of the race dynamics. By combining EDA, predictive modeling and statistical analysis, this research offers insights for improving race strategies, optimizing pit stops and increasing team competitiveness. It also contributes to the growing use of data analytics in motorsport, demonstrating how advanced analytical methods can be used for more effective strategic planning.**

***Keywords*—strategy, Formula 1, EDA, regression models**

I. Introduction

Formula 1 strategy optimization has become an increasingly data-driven research field. Heilmeier et al. (2020) [2] introduced the concept of a Virtual Strategy Engineer (VSE) using artificial neural networks to make real-time strategy decisions, integrating the latest technological advances in the field of sensors and telemetry for a more accurate analysis of vehicle performance in dynamic race conditions. Aguad and Thraves (2024) [3] developed a comprehensive model using dynamic programming and game theory to optimize pit stop strategies, allowing for a better understanding of the complex interaction between different factors in real time. At the same time, Hettmann (2024) [4] showed how machine learning models can predict the optimal pit stop time by analyzing tyre wear and driver characteristics, taking into account specific track conditions. Thomas et al. (2025) [5] introduced reinforcement learning for strategy in Formula 1, combining strategic decision-making with transparency in artificial intelligence. Neumann and Zardini (2024) [6] introduced an innovative approach to co-designing vehicle components and racing strategies, while Quiroga García et al. (2024) [7] developed a comprehensive new approach based on historical data from 1950 to 2023, focusing on the evolution of racing strategies. While these studies have significantly advanced the understanding of race strategy, my research takes a different approach by focusing on Exploratory Data Analysis (EDA) and regression models to identify key factors influencing lap time and position change. Unlike previous studies that primarily simulate strategies, this research aims to provide data-driven insights directly from real-world 2024 season data, ensuring a more adaptable and evidence-based approach to race strategy optimization.

II. Methodology

Formula 1 is a sport of split-second decisions and calculated risks. Understanding winning strategies requires more than intuition – it demands data-driven analysis. Examining key factors like tyre selection, pit stops and weather conditions starts with structured data processing, Exploratory Data Analysis (EDA) and regression modeling. This methodology serves as a systematic approach to uncover insights that refine our understanding of race strategy.

A. Data

This research is based on a comprehensive dataset from the 2024 Formula 1 season, up to and including the Dutch Grand Prix. The data was collected via the FastF1 API, a recognized source for detailed Formula 1 racing data and is organized into four main datasets:

- Lap Data (17,123 rows, 32 columns) – Containing lap times, sector performance, tyre compound and tyre life, enabling analysis of race development.
- Result Data (299 rows, 22 columns) – Including final positions, lap count and completion status for each driver.
- Circuit Data (810 rows, 8 columns) – Covering track characteristics such as length, laps and corner details.
- Weather Data (2,288 rows, 9 columns) – Providing meteorological conditions recorded at one-minute intervals during each race.

These structured datasets provide a rich basis for analyzing various aspects of race strategies across multiple races, drivers, and conditions.

B. Data Preprocessing

Raw Formula 1 race data is complex and requires systematic preprocessing for accurate analysis. We structured this process into the following key stages:

- Data transformation – Converted time formats to numeric values, standardized numeric variables and encoded categorical data.

- Feature creation – Derived key metrics such as pit stop duration and lap time differences, to track performance variations.
- Data integration – Merged datasets while ensuring temporal and logical consistency, linking lap data to final race results for a unified dataset.

C. Analysis techniques

To analyze race performance, we applied multiple analytical models focusing on lap times, position changes, tyre usage and pit stops:

- Exploratory Data Analysis – Conducted descriptive statistics and visualizations (scatter plots, histograms, box plots) to identify trends in lap times, tyre performance and pit stop effects. Correlation analysis provided insights into relationships between key variables.
- Regression Models – Implemented two advanced models to quantify performance factors:
  - Random Forest Regressor – Captured nonlinear interactions and identified key variables affecting lap times and position changes.
  - Gradient Boosting – Improved accuracy by iteratively refining predictions, effectively handling nonlinear race performance trends.

By combining these techniques, we developed a comprehensive strategy analysis, uncovering key performance patterns and evaluating strategies across different scenarios.

III. Analysis and Results

In Formula 1 precision in decision-making is key to success. By analyzing the 2024 season data, we examined how racing strategies shape outcomes. This analysis identifies patterns in lap times, tyre choices and pit stops, revealing their impact on performance.

Each decision, from tyre selection to pit timing, triggers a chain reaction affecting the final standings. Advanced regression models quantify these effects, offering deeper insights into performance factors. The following sections present the key findings, outlining the determinants of success in modern Formula 1.

A. Exploratory Data Analysis

In this chapter, we present the results of a detailed exploratory analysis, focusing on three key aspects of racing events. By examining lap positions and times, tyre selections and pit stop characteristics, we establish a foundation for understanding the factors that influence performance. This analysis uncovers not only clear trends but also subtle patterns that are often overlooked in routine race evaluations.

1) Position and Lap Time Analysis: To get an initial insight into the dynamics of the 2024 season, we started by analyzing the teams' final standings across all races shown in Figure 1. We created a heatmap that shows each team's finishing positions, with teams ranked by championship points. Blue shades indicate higher finishing positions, while red shades indicate lower finishing positions. This visualization reveals not only the differences between the teams, but also how their performance varies from race to race.

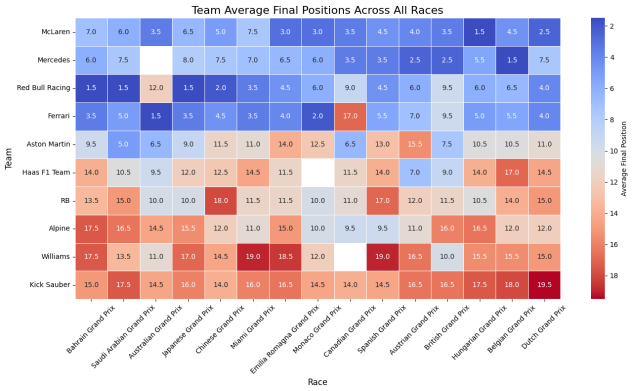


Fig. 1. Team Average Final Positions Across All Races

McLaren, Red Bull, Mercedes and Ferrari dominate with consistently high rankings, while Ferrari's standout performance in Australia (1.5 avg. position) highlights their potential. In the midfield, Aston Martin and Haas show mixed results, with Aston Martin's 7.5 avg. position at Silverstone reflecting track influence. At the lower end, Williams, Alpine, and Kick Sauber rank consistently lower, while RB and Haas struggle at circuits like Miami and the UK, indicating setup or strategy challenges.

Beyond overall team performance, individual lap time analysis reveals key factors influencing race pace.

Figure 2 shows Sector 1 has the highest impact on lap time (0.76 correlation), making it crucial for a competitive lap. Sector 2 has a moderate effect (0.59), while Sector 3 contributes the least (0.40). Teams should prioritize optimizing Sector 1, as early losses are hard to recover later in the lap.

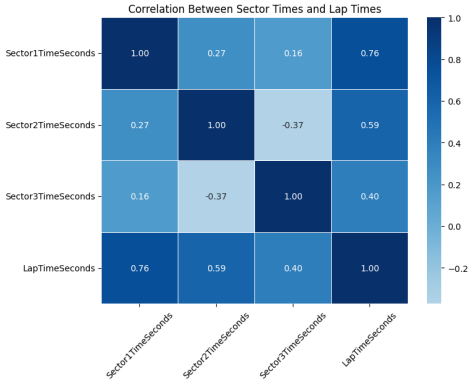


Fig. 2. Correlation Between Sector Times and Lap Times

The next key factor in performance is tyre consistency, measured through lap delta times—the variation in lap times during a stint. Figure 3 visualizes this through a box plot, highlighting differences in tyre stability and degradation.

Tyre selection is one of the most critical strategic decisions in Formula 1, directly impacting race results. As Tremlett & Limebeer (2016) [10] point out, track conditions – temperature, surface grip and circuit layout – have driven tyre manufacturers to develop different compounds.

Teams have three basic types of tyres available for dry conditions – Hard, Medium and Soft, as well as Intermediate and Wet tyres for wet conditions. Each tyre compound has different characteristics in terms of durability, degradation rate and overall speed. Hard tyres last longer but provide less grip, while soft tyres are much faster but wear out more quickly. Intermediate and wet tyres are used in wet conditions, where choosing the right type of tyre at the right time can be crucial for race performance.

The results show that soft and medium tyres maintain stable lap times with low variability, especially in the first stint when they are fresh. In contrast, intermediate and wet tyres exhibit higher variability due to changing track conditions, confirming their unpredictability.

As stints progress, hard and intermediate tyres degrade more noticeably, with increased lap time fluctuations. While soft tyres remain relatively stable, hard tyres show greater performance swings, making them harder to manage in longer stints. Strategic takeaways:

- Soft and medium tyres provide predictable lap times, making them ideal for high-performance stints.
- Hard tyres last longer but become less consistent, requiring careful strategy.
- Intermediate and wet tyres are highly track-dependent, making their performance unpredictable.

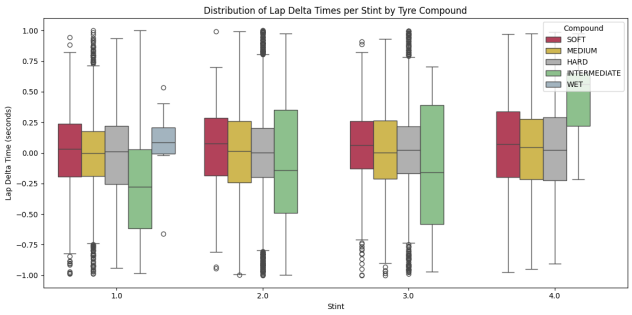


Fig. 3. Distribution of Lap Delta Times per Stint by Tyre Compound

2) Tyre Compounds Analysis: Figure 4 provides a visual representation that reinforces the previous analysis of tyre performance and durability. This box plot highlights the differences in tyre lifespan, confirming the strategic role of tyre selection in Formula 1.

Hard tyres are the most durable, with some stints exceeding 40 laps, highlighting their resistance to thermal and mechanical wear. Medium tyres offer a balance between performance and longevity, typically lasting 15 to 25 laps, making them a flexible choice in variable conditions. Soft tyres have the shortest lifespan, with most stints ranging from 10 to 15 laps, prioritizing grip over endurance for aggressive starts or decisive final laps.

Intermediate and wet tyres show high variability, fully dependent on weather conditions, making their longevity unpredictable. This reinforces the importance of real-time strategy adjustments. Ultimately, success in Formula 1 is no longer just about raw speed but about how well teams and drivers anticipate, adapt and manage tyres across changing conditions.

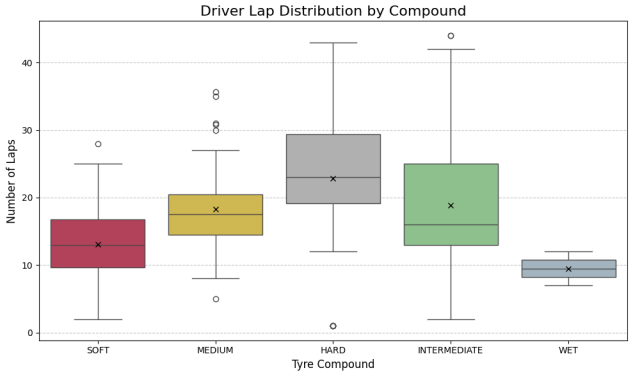


Fig. 4. Driver Lap Distribution by Compound

3) Pit Stop Analysis: Pit stops are more than just tyre changes - they are strategic decisions that impact position, tyre management and overall race performance.

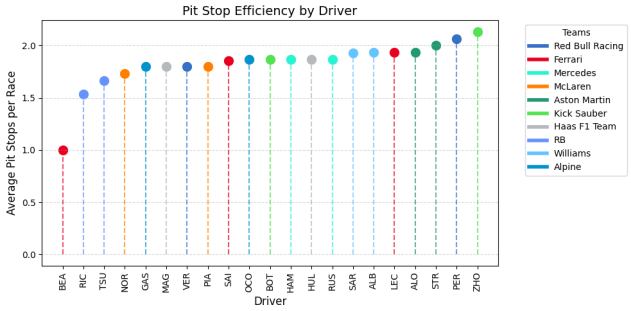


Fig. 5. Average Pit Stops per Race by Driver and Team

The number of stops depends on several factors:

- Tyre degradation – Faster wear leads to more frequent stops.
- Weather conditions – Sudden changes require tyre switches.
- Track incidents – Safety cars or damage alter strategies.
- Driver approach – Aggressive drivers often pit more to gain positions.

Figure 5 highlights key patterns:

- Fewest pit stops (1) – Focus on tyre preservation, minimizing lost time and efficient stint planning.
- Most pit stops (2+) – Favor aggressive strategies, fresh tyres and higher risk for gains.
- Teammate variations – One driver may prioritize position, while the other takes risks.

This analysis confirms that modern Formula 1 success depends not just on speed, but on adaptability. Pit stops are strategic chess moves that can redefine a race.

B. Connections between race factors

In Formula 1 no element exists in isolation. Every aspect of a race – from tyre selection to pit stops – forms part of a complex, interconnected system. This dynamic interplay is what makes motorsport both fascinating and unpredictable.

Tyres are more than just rubber, they are a key strategic tool influencing lap speed, endurance and overall race tactics. Their management affects pit stop frequency, performance timing and degradation control. External factors like track temperature, humidity and rain further complicate decisions, forcing teams to balance aggression with consistency. In such scenarios, real-time data analysis is just as crucial as car performance.

1) Tyre Compound vs Lap Time: Figure 6 reveals a complex relationship between tyre compounds and lap times. While softer tyres traditionally offer better grip and speed, data shows that medium tyres have the fastest average lap time (86.0s), followed by hard tyres (86.8s), with soft tyres slightly slower (88.1s). Intermediate (92.8s) and wet (95.5s) tyres are the slowest due to higher drag and lower grip.

Soft tyres excel in qualifying but degrade quickly in races, especially when used in high-fuel conditions. Medium and hard tyres are preferred for longer stints, optimizing race pace and reducing pit stops. While soft tyres enable aggressive strategies, their rapid wear impacts overall performance. Track and weather conditions further influence tyre behavior, with some circuits favoring medium compounds. Contrary to expectations, this analysis confirms that medium and hard tyres often provide the most stable and competitive race pace.

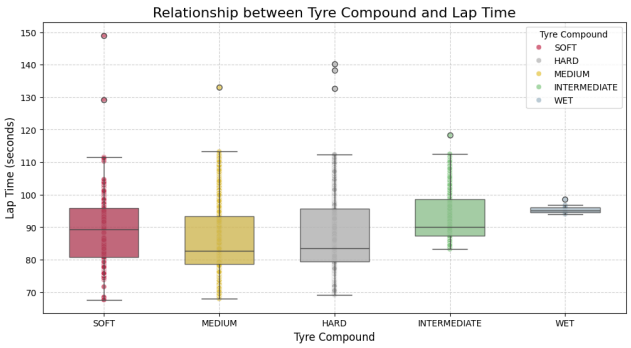


Fig. 6. Relationship between Tyre Compound and Lap Time

2) Starting Tyre Type vs Final Classification: The analysis of starting tyres and their impact on final classification reveals key strategic choices at the race start. Figure 7 shows a comparison of tyre compounds and podium finishes, highlighting clear patterns.

Most podiums (35) came from drivers starting on medium tyres, indicating their strong balance between performance and durability. In contrast, soft tyres, despite being the fastest per lap, resulted in only 6 podiums, likely due to higher degradation forcing early pit stops. Intermediate tyres saw 3 podiums, reflecting races with variable weather, while hard and wet tyres led to none, suggesting they are rarely optimal for podium contention.

This analysis confirms that starting tyre choice plays a crucial role in shaping race outcomes.

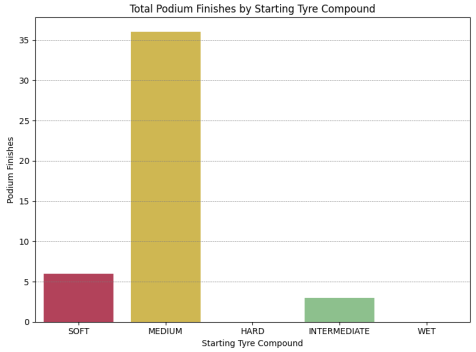


Fig. 7. Total Podium Finishes by Starting Tyre Compound

3) Finishing Tyre Type vs Final Classification: This analysis examines the impact of the final tyre type on race results, following the same approach as the starting tyre analysis. Figure 8 shows that hard tyres secured most podiums (31), emphasizing their durability in the final stages. Soft and medium tyres had around seven podiums each, likely used for specific strategies. Intermediate and wet tyres had none, as teams switch to dry compounds when conditions improve. The most effective strategy is starting on medium and finishing on hard tyres.

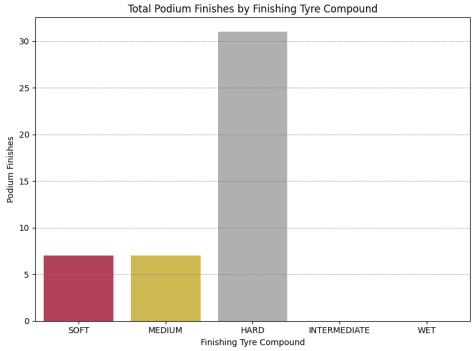


Fig. 8. Total Podium Finishes by Finishing Tyre Compound

4) Number of stops vs Final Position: Figure 9 highlights the relationship between pit stops and final positions, offering a data-driven approach to optimal race strategies. The heat map shows that:

- One-stop strategies work best at tracks like Monaco (avg. 6.5), where track position is key.
- Two stops are ideal in Canada (avg. 6.9) and China (avg. 7.8), while three stops lead to worse results.

- Three stops at Belgium (avg. 7.0) performed slightly better than two (avg. 8.8), suggesting some tracks benefit from extra stops due to high degradation.

This analysis confirms that pit stop strategies must be race-specific, with factors like tyre wear, overtaking difficulty and track characteristics shaping the best approach.

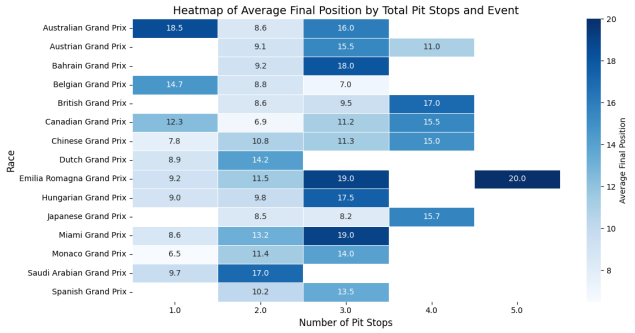


Fig. 9. Heatmap of Average Final Position by Total Pit Stops and Event

5) Race Length vs Strategy: This analysis, on Figure 10, categorizes race length by lap count, not track length, classifying races as short, medium or long. Short races, covering more kilometers, tend to have higher and more variable pit stops due to increased tyre degradation and race interruptions. Medium and long races show fewer, more consistent pit stops, as teams prioritize track position.

Lap times reveal an inverse trend – short races have the longest lap times since they take place on longer tracks, while long races have the shortest laps due to more compact circuits.

Tyre selection follows clear patterns. Hard tyres dominate across all race lengths, minimizing pit stops. Medium tyres serve as a balanced alternative, while soft tyres are favored in medium-length races for aggressive stints. Intermediate and wet tyres remain condition-dependent. This confirms that race length significantly shapes strategy, requiring teams to optimize pit stops and tyre choices accordingly.

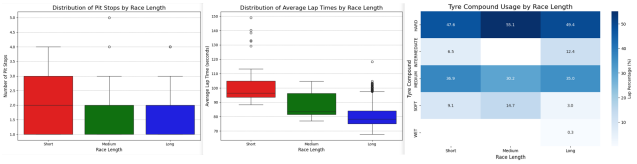


Fig. 10. Left: Distribution of pit stops by race length, Center: Distribution of average lap times by race length, Right: Heatmap - percentage usage of tyre type by race length

C. Regression models

In Formula 1 strategic decisions depend on multiple interrelated factors, from tyre choice and pit stops to weather and track characteristics. While previous analyses identified trends, the next step is to quantify their impact on race outcomes.

To achieve this, we used regression models to assess how different variables influence race performance, focusing on lap times and position changes. By applying Random Forest Regressor and Gradient Boosting models, this approach assigns weight to each factor, enabling a data-driven optimization of race strategies.

1) Lap Time Analysis: We used a Random Forest Regressor for each race separately to account for track-specific characteristics, such as surface conditions and tyre degradation. Modeling each race individually provided better predictive performance than a single model for all races.

Figure 11 presents the model's R^2 Score and Mean Squared Error (MSE) across races. High R^2 values confirm that the model effectively explains lap time variations, while higher MSE in some races indicates greater unpredictability due to external factors like safety cars and weather conditions.

The model also identifies the most significant factors affecting lap times:

- Lap number has the highest importance (above 0.8), reflecting tyre wear, fuel depletion and race progression.
- Starting position strongly influences lap times (above 0.5), especially in the early race stages.
- Sector speeds, particularly in sector two, significantly impact lap times.
- Track temperature, pit stops and tyre life also contribute, reinforcing the role of strategy.

Random Forest was chosen for its ability to handle nonlinear relationships, resist overfitting, rank feature importance and adapt to different tracks. Using 200 decision trees provided an optimal balance between accuracy and efficiency, ensuring stable predictions without excessive computational cost.

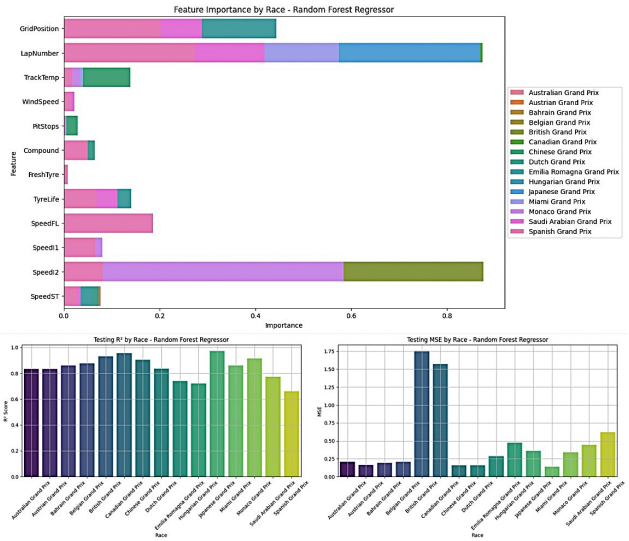


Fig. 11. Random Forest Regressor for Lap Time analysis, showing feature importance, R^2 scores, and MSE across races

Gradient Boosting, unlike Random Forest, builds trees sequentially, improving accuracy by correcting errors at each step. We used it to analyze sector times' impact on lap time, identifying their incremental contributions.

Figure 12 shows the model's accuracy in predicting lap times and highlights that Sector 1 has the greatest influence, followed by Sector 2, with Sector 3 having the least impact. This aligns with Figure 2, which also showed a stronger correlation between Sector 1 and lap time.

Sector 1's dominance is likely due to its long straights, braking zones and fast corners, setting the pace for the lap. Optimizing braking, corner exits and tyre management in this sector is crucial for improving overall performance.

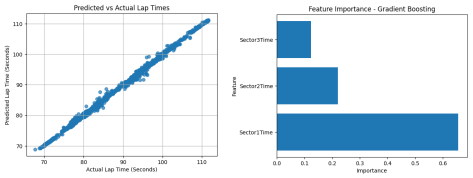


Fig. 12. Gradient Boosting Regressor for Lap Time analysis, showing actual vs. predicted times and feature importance by sector

2) Position Change Analysis: To analyze position changes, we trained a Random Forest Regressor model for each race separately to capture track-specific factors. The results, shown in Figure 13, confirm that starting position is the most significant factor, reinforcing the importance of qualifying. Drivers who start at the front have a much greater chance of maintaining or improving their position, while those further back struggle to overtake.

Pit stops also play a key role, with strategy significantly impacting final positions. Faster sector times indicate better race pace, while wind speed and track temperature have a minor influence. The model performs well, confirming that race-specific factors shape position changes, with strategy and track position being the most decisive elements.

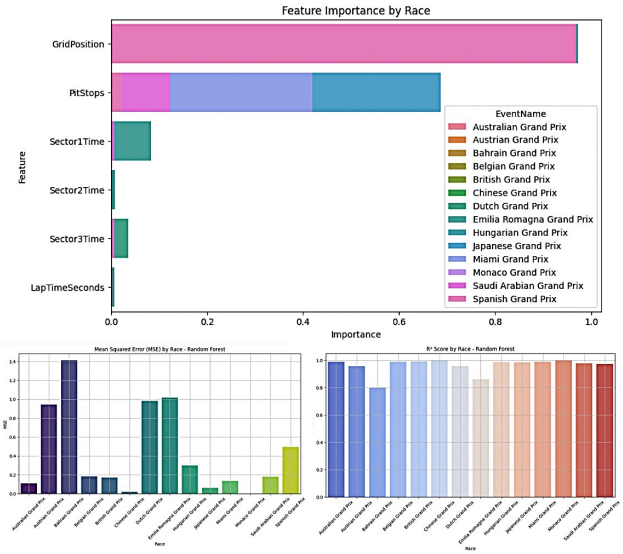


Fig. 13. Random Forest Regressor for Position Change analysis, showing feature importance, R^2 scores, and MSE across races

Conclusion

Optimizing race strategy in Formula 1 depends on several key factors, each playing a crucial role in determining the final outcome.

- Starting Position: The initial grid position has the most significant impact on the final classification, emphasizing the importance of strong qualifying performance.
- Pit Stop Strategy: Properly timed pit stops can be the difference between victory and defeat, making strategic decision-making critical.
- Tyre Selection: Medium and hard tyres generally lead to better race results, highlighting the importance of choosing the right compound based on race conditions.
- Track Characteristics: The first sector is particularly important, as it often dictates the flow of the rest of the race.

To optimize race strategies, teams must focus on two essential aspects:

1) Understanding the Factors – A deeper understanding of racing variables can lead to more efficient strategies.
2) Adapting Strategies – Teams must tailor their approach based on the unique characteristics of each track to maximize performance.

By integrating these elements, teams can develop more effective strategies, improving their chances of success in highly competitive racing environments.

References

[1] Tukey, John W. 1977. Exploratory Data Analysis. Reading, MA: Addison-Wesley Publishing Company.
[2] Heilmeier, Alexander, Andreas Thomaser, Maximilian Graf, and Johannes Betz. 2020. "Virtual Strategy Engineer: Using Artificial Neural Networks for Making Race Strategy Decisions in Circuit Motorsport." Applied Sciences 10 (21): 7805.
[3] Aguad, Felipe, and Cristian Thraves. 2024. "Optimizing Pit Stop Strategies in Formula 1 with Dynamic Programming and Game Theory." European Journal of Operational Research.
[4] Hettmann, Vincent. 2024. "From Data to Podium: A Machine Learning Model for Predicting Formula 1 Pit Stop Timing." Doctoral dissertation, Universidade NOVA de Lisboa.
[5] Thomas, Daniel, Jiarui Jiang, Avinash Kori, Alessandro Russo, Stefan Winkler, Simon Sale, James McMillan, and Francesco Belardinelli. 2025. "Explainable Reinforcement Learning for Formula One Race Strategy." arXiv preprint arXiv:2501.04068.
[6] Neumann, Maximilian P., and Gioele Zardini. 2024. "On the Co-Design of Components and Racing Strategies in Formula 1." In IEEE Intelligent Vehicles Symposium (IV 2024). ETH Zurich.
[7] Quiroga García, María Paz. 2024. "Optimization of Pit Stop Strategies in Formula 1 Racing: A Data-Driven Approach." Master's thesis, Universidad de Chile.
[8] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." Journal of Machine Learning Research 12: 2825–30. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.
[9] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." Journal of Machine Learning Research 12: 2825–30. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.
[10] Tremlett, A. J., and D. J. N. Limebeer. 2016. "Optimal Tyre Usage for a Formula One Car." Vehicle System Dynamics 54 (10): 1448–73.

AI-Powered Pipeline for Video Generation from Large Texts

1st Andrej Nankov
MSc Candidate, Software Engineer
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, N.Macedonia
andrej.nankov@students.finki.ukim.mk

2th Vladimir Trajkovik
PhD, professor
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, N.Macedonia
trvlado@finki.ukim.mk

***Abstract*—The demand for video content is growing fast, driven by AI advancements, the high-speed Internet, and monetization on platforms such as YouTube and TikTok. Generating high-quality videos from large textual formats, such as books and long-form articles, remains challenging due to contextual inconsistencies and information loss. This paper presents a structured AI-powered processing pipeline designed to transform extensive textual content into video materials while preserving coherence and key information. The pipeline consists of four key stages: (1) Text Summarization and Keyword Extraction, utilizing the TextRank algorithm to generate concise and meaningful content; (2) Configuration Planning, where video elements are structured in JSON format for seamless AI-based processing; (3) AI-driven content Generation, integrating ChatGPT for script development and InvideoAI for automated video production; and (4) Post-processing and User Refinement, ensuring adaptability and editability of the final output. The proposed method is evaluated in practice using both article-based and book-chapter inputs, demonstrating its effectiveness in generating contextually relevant and structured video content. Experimental results indicate that the pipeline successfully reduces manual intervention and improves processing efficiency, though challenges remain in handling non-textual elements such as images, graphs, and formulas. Future work will explore enhanced AI summarization techniques, multi-modal content integration, and real-time video generation improvements. This research contributes to the growing field of AI-assisted content creation by providing a scalable and adaptable approach to transforming large-scale textual data into engaging video materials.**

***Keywords*—AI, Text Summarization, Video Processing Pipeline, Content Generation, Media Transformation.**

I. Introduction

In the era of digital transformation, the demand for automated content generation increases driven by advances in artificial intelligence (AI), natural language processing (NLP), and video production technologies. The ability to convert large-scale textual content, such as books, research papers, and long-form articles, into engaging video material has significant applications in education, media, and digital marketing. Existing AI-driven video generation tools often struggle with context retention, coherence, and scalability [1] [2], which leads to the need for a more structured and efficient processing pipeline. Advancements in artificial intelligence (AI) have led to the rapid development of tools such as Microsoft Copilot, Google Gemini, Meta AI, and ChatGPT, which are widely used for reasoning-based tasks [3]. Among recent AI-based tools, DeepSeek has introduced DeepSeek-R1 [4], a reasoning model built using reinforcement learning (RL) without supervised fine-tuning. This model demonstrates remarkable reasoning abilities in benchmark evaluations, effectively addressing challenges such as poor readability and language inconsistencies. A key advantage of DeepSeek-R1 is its ability to operate on lower computing power, making it suitable for self-hosted applications and expanding the possibilities for AI deployment. Currently, DeepSeek-R1 processes only textual data, but it holds significant potential for accelerating AI-driven content processing pipelines, particularly for tasks involving large-scale text analysis and transformation. Benchmark results indicate that DeepSeek-R1 achieves performance levels comparable to premium models like ChatGPT-4o [4], making it an attractive alternative for cost-effective AI integration. While ChatGPT remains advantageous due to its easy API integration and affordability under subscription-based models, DeepSeek-R1 offers greater flexibility for self-hosted AI solutions [5]. This research presents an AI-powered workflow designed to transform extensive textual data into structured video narratives, leveraging text summarization, keyword extraction, structured configuration planning, and AI-assisted video synthesis. By addressing the challenges of automated text-to-video conversion, this study contributes to enhancing digital content accessibility, optimizing AI-driven storytelling, and improving the scalability of video content production. This paper is structured as follows: Section II provides background information for the paper. Section III reviews existing data processing frameworks discussing their relevance to AI-driven video creation. Section IV details the proposed processing pipeline, outlining its core stages. Section V presents experimental results, evaluates the pipeline and performance using articles and book chapters, and discusses key challenges. Section VI proposes potential enhancements in multi-modal AI processing and real-time video generation. Finally, Section VII concludes the paper.

II. Background

This research is inspired by two key previous projects: a multi-tenant e-learning platform and a SaaS-based video edit-

ing platform. Both projects are fully developed, production-ready solutions that have provided valuable insights into scalability, automation, and processing pipelines, which form the foundation of this study.

A. Multi-Tenant E-Learning Platform

One of the biggest challenges in e-learning platforms is ensuring efficient video availability for course materials. The proposed solution incorporates RabbitMQ Stream, decentralized object storage, and Content Delivery Networks (CDN) to enhance video distribution, scalability, and system responsiveness. The multi-tenant architecture ensures that multiple users and organizations can access the platform simultaneously while maintaining performance and data isolation. This experience has shaped the research by emphasizing the need for automated, scalable, and flexible video content management.

B. SaaS-Based Video Editing Platform

The second project focuses on video editing in a SaaS environment, where the video timeline creation process follows a structured pipeline. Each stage in this pipeline is responsible for a specific function, ensuring an organized workflow for video processing and rendering. The platform supports essential editing actions such as merging, cutting, muting, ordering, and drag-and-drop operations.

A key challenge on this platform was to handle longer video formats while maintaining efficient execution. To address this, Laravel Horizon was integrated, enabling asynchronous execution, job queueing, and real-time performance monitoring within the Laravel ecosystem. This system played a crucial role in understanding the complexities of video processing pipelines, which directly influenced the research presented in this paper.

C. Data Pipelines and Their Role in AI Processing

Data pipelines play a fundamental role in automating data transformation, enrichment, and processing before applying AI models. A typical pipeline consists of:

1) A Source - the raw input data, such as text, video, or structured datasets.
2) Processing Steps - A series of transformations, such as filtering, grouping, summarization, and AI-based analysis.
3) A Destination - The final structured output, ready for storage or further application.

For example, a single social media comment can be processed through different pipelines for real-time monitoring, sentiment analysis, or location-based mapping. Each application uses a distinct pipeline that must function seamlessly and efficiently.

Pipelines are widely used in event-driven architectures, enabling real-time processing while optimizing computing costs. The approach used in this research follows a batch processing model, where large textual data are processed asynchronously in stages. This method ensures scalability while allowing AI models to generate contextually accurate and structured video content.

III. Related work

Traditional data processing frameworks, such as MapReduce and Spark, have transformed the way large-scale textual and semi-structured data are handled [6]. However, video processing presents unique challenges due to the complexity and size of video files. Addressing these challenges requires scalable and efficient solutions that optimize resource usage while maintaining low latency and high processing accuracy.

One tool designed to tackle these challenges is Sprocket [6], a serverless video processing framework that aims to push the boundaries of cloud-based video transformation by offering an optimized, low-cost solution. Unlike text-based data pipelines, Sprocket is specifically designed to process video files and operates on a serverless cloud infrastructure.

Sprocket follows a modular pipeline model, breaking video processing tasks into distinct stages, such as decoding, filtering, and encoding. The framework leverages AWS Lambda to achieve dynamic scaling and efficient resource allocation. By distributing tasks across thousands of Lambda instances, Sprocket achieves high parallelism and low latency. The core benefits of Sprocket include:

1) Modularity - Developers can customize processing pipelines by composing multiple stages.
2) Scalability - The serverless architecture enables efficient handling of large-scale video processing tasks.
3) Cost-Effectiveness - The pay-as-you-go model minimizes costs while maintaining high-performance efficiency.
4) AI Integration – Sprocket can incorporate AI-based video analysis tools, such as Amazon Rekognition, to perform advanced tasks such as facial recognition and object detection.

Another notable example of data pipeline optimization is LinkedIn's real-time activity data processing system [7]. Managing real-time data streams presents significant challenges, particularly in advertising, search, recommendation systems, and security. LinkedIn's real-time pipeline is built around Apache Kafka, handling over 10 billion message writes daily across multiple subscribing systems.

Initially, LinkedIn relied on separate batch-oriented systems for user activity data and server metrics, which lacked real-time data access [7]. The inefficiencies in these legacy systems led to difficulties in correlating operational issues with business metrics. While LinkedIn experimented with off-the-shelf messaging solutions like ActiveMQ, performance issues under production loads led to the development of Kafka, a high-throughput, low-overhead data infrastructure.

Kafka functions as a write-ahead log, allowing subscribers to read and process data changes efficiently. Its design principles include:

- Data Partitioning - Distributing data across multiple servers to enhance scalability.
- Batch Processing and Data Compression - Improving performance while reducing network and file system overhead.

- Schema Management - Using Avro for data serialization to ensure compatibility and integration across multiple applications.

Kafka supports over 10 billion daily message writes, with sustained throughput exceeding 172,000 messages per second across 367 topics [7]. LinkedIn integrates Kafka with Hadoop, automatically loading data and structuring it into Hive tables for efficient batch processing [7].

A. Security Considerations in AI-Powered Pipelines

Ensuring the security and integrity of AI-driven pipelines is crucial, particularly when handling sensitive data. Vulnerabilities in AI models, such as data exposure, adversarial manipulation, and privacy risks, must be addressed [8]. Access control mechanisms, data encryption, and continuous monitoring are necessary to protect against security threats [8]. When integrating general AI models, it is essential to prefilter sensitive data before sharing it with external AI tools. Additionally, ensuring compliance with data protection laws and AI ethics guidelines is necessary to prevent misuse or unauthorized data access [9].

IV. Processing pipeline framework

The proposed processing pipeline (Fig. 1) follows the divide-and-conquer principle, where a long textual input (e.g., a book or article) is broken into manageable sections. Each section undergoes a structured pipeline, ensuring that essential content is extracted, transformed, and converted into video format while maintaining coherence and relevance. This modular approach allows efficient processing, where multiple short videos can be generated from different sections and later merged into a cohesive final video product. This strategy is beneficial for individual content creators, businesses, and educational institutions that require structured and contextually accurate video content while minimizing manual editing efforts.

The core pillars of this research are Data, AI, Transformation, and Execution, each corresponding to a key stage in the pipeline. The workflow consists of the following stages:

A. Initialization

To implement the processing pipeline, we employed Node.js along with a selection of built-in and third-party packages. Node.js was chosen due to its scalability, extensive community support, and widespread adoption in modern web and backend development. The pipeline operates primarily in a console-based environment, leveraging the built-in file system module to run multiple pipelines in parallel, thereby optimizing the time required to generate multiple videos per book or document. This stage is responsible for preparing the initial data for the next stages. For instance, read from a file, prepare the text from the defined sections, and remove unused empty spaces and strange characters over the text.

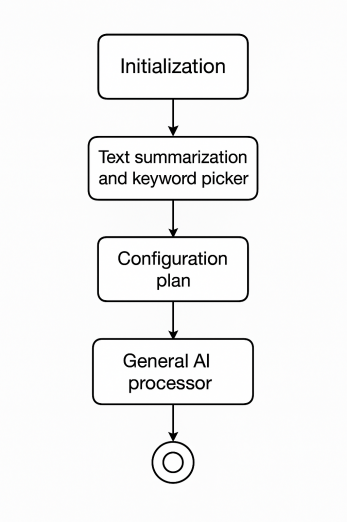


Fig. 1. Processing pipeline stages

B. Text Summarization and Keyword Extraction

The second stage involves text summarization [10] and keyword extraction using a node summarizer, which implements the TextRank algorithm. This algorithm is particularly effective for summarizing text from books, newspaper articles, and research papers by identifying key sentences and ranking them based on their relevance [11].

Summarization Process:

1) Sentence Tokenization – The text is split into individual sentences.
2) Preprocessing - Punctuation is removed, and all text is converted to lowercase.
3) Keyword Mapping - Each sentence is mapped to an array of nouns and adjectives.
   - Example: ”The San Antonio is a good basketball team.” will be taken as [San Antonio, good, basketball, team].
4) Graph-Based Ranking - A weighted graph is constructed where:
   - Each node represents a sentence.
   - Edges are formed between sentences based on shared nouns and adjectives.
   - Edge weights are determined by the frequency of shared words.
5) Random Walk Algorithm - Sentences are iteratively ranked based on their connectivity, with higher-ranked sentences forming the final summary.

This process ensures that the generated summaries retain essential information while eliminating redundancy.

C. Configuration planning

The third stage defines the video structure and behavior using JSON-based configurations. JSON was selected due to its lightweight nature, human readability, and compatibility

```
{
  "title": "Habit Formation: The Power of Repetition",
  "description": "Exploring the transformative process of habit formation through repetition and practice.",
  "scenes": [
    {
      "scene_id": 1,
      "visuals": {
        "type": "image",
        "source": "Illustration of neural pathways forming with repetition"
      },
      "narration": "Habit formation is a remarkable journey where behaviors transition from conscious actions to automatic responses through repetitive practice."
    },
    {
      "scene_id": 2,
      "visuals": {
        "type": "text_overlay",
        "elements": [
          "Automaticity",
          "Repetition",
          "Behavior Change"
        ],
        "background_type": "abstract",
        "background_color": "green"
      },
      "narration": "The duration to solidify a habit varies; what truly matters is the consistent repetition that engrains the behavior into our daily routines."
    },
    {
      "scene_id": 3,
      "visuals": {
        "type": "text_list",
        "elements": [
          "Practice",
          "Consistency",
          "Transformation"
        ],
        "background_type": "minimal"
      },
      "narration": "Building a habit is less about the time it takes and more about the dedication to practice. With persistence, habits evolve from intentional efforts to subconscious actions."
    },
    {
      "scene_id": 4,
      "visuals": {
        "type": "image_collage",
        "images": [
          "Image of a person repeating a behavior",
          "Brain scan showing habit formation",
          "Illustration of habit becoming automatic"
        ]
      },
      "narration": "Repetition is key in habit formation. Through consistent practice, behaviors shift from requiring attention to seamlessly integrating into our daily lives."
    },
    {
      "scene_id": 5,
      "visuals": {
        "type": "text_overlay",
        "elements": [
          "Habit Formation",
          "Practice",
          "Consistency"
        ],
        "background_type": "mindful_patterns"
      },
      "narration": "Embrace the journey of habit formation. Explore the power of repetition, practice, and unwavering consistency in transforming behaviors into automatic habits."
    }
  ],
  "audio": {
    "background_music": "upbeat_and_professional",
    "narration_voice": {
      "gender": "male",
      "tone": "friendly",
      "accent": "neutral"
    }
  }
}
```


Fig. 2. Example JSON configuration

with multiple programming languages, making it ideal for structured data exchange.

The configuration file specifies:

1) Video metadata (e.g., title, description).
2) Scene elements, including:
   - Type of visual effects.
   - Selection of backgrounds and colors.
   - Image overlays for enhanced representation.
   - A supportive narrator to provide voiceovers.
3) Audio parameters:
   - Background music selection.
   - Narrator attributes (tone, accent, gender).

Once finalized, this configuration is attached to the summarized text and passed to the AI-powered video generation stage. See (Fig. 2) for an example JSON configuration.

D. Configuration planning

The final stage leverages AI models to convert the processed content into fully structured video material. This is achieved through two API-based interactions:

1) ChatGPT API (GPT-3.5-Turbo, GPT-4-Turbo) – Generates a cohesive and engaging narration.
2) InvideoAI API – Automates video creation and rendering, integrating text, images, and narration into a structured video format.

Once the request is processed, the InvideoAI dashboard provides:

- A progress bar indicating the rendering status.
- Options for previewing and editing specific sections.
- Download capabilities for high-resolution video output (1080p at 60fps).

Through this approach, users gain greater control and flexibility, ensuring high-quality, AI-generated video content without losing contextual accuracy.

V. Evaluation of the Processing Pipeline

The processing pipeline was tested using different articles and book chapters to evaluate its effectiveness in generating contextually relevant videos. We can summarize that (1) The initial stage operated almost identically over different long texts according to time-consummation and processing power. (2) The text summarization and keywording stage is computationally lightweight, making it ideal for summarization and there is no need for pre-trained data. (3) The configuration plan directly depends on what we added to the configuration JSON format. Usually, this configuration is persistent but it offers flexibility to make the changes before we jump to the next stage. (4) The general AI processor is the most time-consuming stage since depends on making HTTP calls and communicating first with the API of ChatGPT and then InvideoAI. This stage is also subscription-based and depends directly on what type of plans we have picked for the ChatGPT model and InvideoAI plan. By opting for higher-tier plans, we can enhance video quality and the AI leverages the training data. Additionally, premium plans eliminate any restrictions on video length within a single pipeline run. But this pipeline can be optimized by integrating agent-based orchestration to enable adaptive task refinement presented in AI Scientist-v2[12], demonstrates the effectiveness of agentic pipelines in automating complex multi-stage tasks.

A. Video Generation from Article

InvideoAI played a crucial role in enhancing video quality by utilizing its built-in keyword-based video storage, which allowed seamless integration of abstract video elements into the generated content. The configuration stage (Fig. 2) performed exceptionally well, enabling automated scene selection based on summarized text, keywords, narration, and background audio metadata. One of the test articles focused on the software industry downturn related to the global economy, COVID-19, and war [13]. The pipeline successfully processed this content through all stages:

- Stage 1 (Initialization): The text was structured and preprocessed.
- Stage 2 (Summarization and Keyword Extraction): The TextRank algorithm effectively identified key terms,

including economy, COVID-19, and war, forming a weighted graph to ensure contextual relevance.
- Stage 3 (Configuration Planning): No modifications were required in this stage, as the predefined configuration effectively structured the video.
- Stage 4 (AI-Powered Video Generation):
  - ChatGPT generated narration scripts based on extracted keywords.
  - InvideoAI produced over 20 video scenes, each featuring distinct backgrounds, keywords, and visual representations (e.g., green screens, minimalistic visuals, and source explanations).
  - The background music was configured to be dramatic with a positive ending, aligning with the article's theme.
  - The narrator was male with a deep English voice, as per the content creator's preferences.

The final video had a duration of 2 minutes and 41 seconds [14] and was generated using the InvideoAI free plan. The tone and structure were preserved, ensuring that the generated content accurately represented the intent of the article.

B. Video Generation from Book

The second test involved summarizing a book chapter from Atomic Habits by James Clear [15]. The selected chapter, "Walk Slowly, but Never Backward", discusses the long-term impact of habit formation and the power of small changes over time.

- Stage 1 and Stage 2 (Initialization and Summarization):
  - The summarization model effectively extracted key takeaways, focusing on the role of small improvements in long-term success.
  - A graph image within the chapter was ignored, as it contained non-textual data. (This limitation is discussed further in the Domain Challenges section.).
- Stage 3 (Configuration Planning):
  - The same configuration template as used for articles was applied, ensuring consistency in narration, visuals, and background audio.
- Stage 4 (AI-Powered Video Generation):
  - The AI extended scene durations, adding more descriptive narration and additional visual effects to illustrate key concepts.
  - Selected images represented brain connections, behavioral repetition, and habit formation to complement the narration.
  - The background music was calming and study-oriented to match the book's self-improvement theme.
  - The narrator was male, with a friendly tone and a standard English accent.

The final video was 52 seconds [16], shorter than the article-based video but focused on delivering precise, structured content. While all textual elements were well-covered, the AI was unable to process and represent numerical data from graphs, indicating a potential limitation in handling non-textual elements.

C. Challenges and Limitations

While the proposed pipeline demonstrates scalability, several challenges must be addressed to enhance performance, cost efficiency, and adaptability.

One major limitation concerns scalability and performance bottlenecks. Although the system supports parallel execution of multiple processing pipelines, increasing throughput by adding more Node.js workers introduces computational complexity. Currently, the pipeline is limited to processing one long text at a time, and when multiple video creations are queued in InvideoAI, delays occur regarding asynchronous processing and load complexity, affecting real-time InvideoAI actions. A potential solution involves implementing a new pipeline stage before the AI processor stage which has dynamic worker allocation based on load-balancing techniques, which covers the asynchronous processing and optimizes resources for load complexity. For example, Pipeflow can be added as a stage to solve this problem [17].

Another critical challenge is the cost of cloud computing. Running the pipeline on cloud-based servers requires significant computational resources, and as the number of workers increases to enhance processing speed, operational costs rise. To address this, cost-efficient cloud strategies, such as serverless computing and AI model quantization, should be explored to reduce expenses while maintaining performance.

A notable limitation is the inability to process non-textual data, such as images, graphs, and mathematical formulas. Since the pipeline primarily relies on text-based transformations, it fails to interpret visual content, leading to information loss, especially in academic materials. A potential solution involves integrating image-processing libraries like OpenCV or Tesseract or using multimodal AI models capable of jointly analyzing text and visual data, which would enhance the system's ability to extract and summarize graphical content [18]. Efficiency in AI processing time is another area that requires improvement. While the preprocessing of multiple book chapters takes less than three seconds demonstrating high efficiency in summarization and keyword extraction, the most significant delay occurs during API communication with ChatGPT and InvideoAI, particularly during scene rendering and video generation. One way to mitigate this is by deploying local AI models, such as running open-source AI models on-premises instead of using ChatGPT models, which would reduce dependency on external APIs, lower latency, and improve overall system responsiveness [19].

VI. Future improvements and research directions

To further enhance the pipeline's usability and effectiveness, several areas of future research can be proposed. Optimizing the data pipeline for video merging would allow multiple generated videos to be combined into a single long-form output, such as full-length educational content, or tuning AI models, such as GPT-3.5 or GPT-4, for domain-specific summarization

could enhance content relevance and reduce unnecessary information, with the possible incorporation of knowledge-graph-based summarization techniques for handling complex subjects [20]. Advancing the pipeline's ability to process both text and images is another key area for improvement. By integrating Vision Transformers (ViTs)[21] and OCR-based solutions[22], the system could interpret charts, graphs, and mathematical formulas, addressing one of the primary limitations identified in the current approach. Moreover, enhancing AI-based speech synthesis by incorporating fine-tuned speech models would allow for adaptive tone modulation based on content type, making the generated videos more engaging and suitable for various applications, including academic content, storytelling, and marketing materials [23]. Cost optimization strategies must be explored to make the pipeline more sustainable. Implementing low-latency alternatives to commercial AI services—such as local inference models or open-source AI alternatives-could significantly reduce reliance on costly third-party APIs while maintaining high-quality video generation.

VII. Conclusion

This paper presents an AI-powered processing pipeline designed to transform large textual content into structured video materials. Using text summarization, keyword extraction, structured configuration planning, and AI-driven video generation, the proposed pipeline converts books and long-form articles into contextually accurate videos while minimizing manual intervention. Experimental results indicate that the pipeline produces structured video outputs that align with the original textual intent. The use of ChatGPT and InvideoAI ensures natural language processing capabilities, allowing cohesive narration and visual storytelling. Limitations in handling non-textual data, such as images, graphs, and mathematical formulas, were observed requiring further enhancements for multimodal content processing. Future research will focus on optimizing the AI models for improved content relevance and summarization accuracy, integrating Vision Transformers (ViTs) and OCR-based solutions to process non-textual elements, and developing cost-efficient AI alternatives to reduce reliance on commercial APIs. Additionally, advancements in speech synthesis and AI-driven tone modulation will further improve engagement and adaptability for different audiences.

References

[1] F. Bao, C. Xiang, G. Yue, G. He, H. Zhu, K. Zheng, M. Zhao, S. Liu, Y. Wang, and J. Zhu, "Vidu: a Highly Consistent, Dynamic and Skilled Text-to-Video Generator with Diffusion Models," arXiv, vol. 2405, no. 04233, pp. 1-45, May 2024. doi: 10.48550/arXiv.2405.04233.

[2] F. Waseem and M. Shahzad, "Video Is Worth a Thousand Images: Exploring the Latest Trends in Long Video Generation," arXiv, vol. 2412, no. 18688, pp. 1-35, Dec. 2024. doi: 10.48550/arXiv.2412.18688.

[3] F. Rosén and M. Rübner, "Augmented Large Language Models for Software Engineering," Master Thesis, Dept. of Design Sciences, Lund University, Lund, Sweden, 2024.

[4] DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv:2501.12948v1 [cs.CL], Jan. 2025.

[5] I. Solaiman, R. Bommasani, D. Hendrycks, A. Herbert-Voss, Y. Jernite, A. Skowron, and A. Trask, "Beyond Release: Access Considerations for Generative AI Systems," arXiv:2502.16701v1 [cs.CY], Feb. 2025.

[6] L. Ao, L. Izhikevich, G. M. Voelker, and G. Porter, "Sprocket: A Serverless Video Processing Framework," in Proceedings of ACM Symposium on Cloud Computing (SoCC '18), Carlsbad, CA, USA, October 11–13, 2018, pp. 263-274. doi: 10.1145/3267809.3267815.

[7] K. Goodhope, J. Koshy, J. Kreps, N. Narkhede, R. Park, J. Rao, and V. Y. Ye, "Building LinkedIn's Real-time Activity Data Pipeline," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2012.

[8] H. I. Kure, P. Sarkar, A. B. Ndanusa, and A. O. Nwajana, "Detecting and Preventing Data Poisoning Attacks on AI Models," arXiv, vol. 2503, no. 09302, pp. 1-9, Mar. 2025. doi: 10.48550/arXiv.2503.09302.

[9] G. Feretzakis, K. Papaspyridis, A. Gkoulalas-Divanis, and V. S. Verykios, "Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review," Information, vol. 15, no. 11, pp. 697, Nov. 2024. doi: 10.3390/info15110697.

[10] A. K. Yadav, A. K. Maurya, Ranvijay, and R. S. Yadav, "Extractive Text Summarization Using Recent Approaches: A Survey," Ingénierie des Systèmes d'Information, vol. 26, no. 1, pp. 109-121, Feb. 2021. doi: 10.18280/isi.260112.

[11] V. Gulati, D. Kumar, D. E. Popescu, and J. D. Hemanth, "Extractive Article Summarization Using Integrated TextRank and BM25+ Algorithm," in Proceedings of the International Conference on Advanced Computing and Communication Systems (ICACCS), 2023, pp. 1-6. doi: 10.1109/ICACCS.2023.123456.

[12] Y. Yamada, R. T. Lange, C. Lu, S. Hu, C. Lu, J. Foerster, J. Clune, and D. Ha, "The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search," arXiv, vol. 2504, no. 08066, pp. 1–18, Apr. 2025. doi: 10.48550/arXiv.2504.08066.

[13] A. Nankov, "Software Industry Depression," Medium, 2023. [Online]. Available: https://medium.com/@nanorocks/software-industry-depression-51c69e0e0ed5. [Accessed: 15-Mar-2025].

[14] A. Nankov, "Video from Article - Software Industry Depression" GitHub repository, Available: https://github.com/nanorocks/created-videos-research-ai-pipeline/blob/main/SoftwareIndustryDepressionArticle.mp4. Accessed: Mar. 15, 2025.

[15] J. Clear, "Walk Slowly, But Never Backward," in Atomic Habits, New York, NY, USA: Penguin Random House, 2018.

[16] A. Nankov, "Video from Book Atomic Habits - Chapter 11" GitHub repository, Available: https://github.com/nanorocks/created-videos-research-ai-pipeline/blob/main/AtomicHabitsChapter11.mp4 Accessed: Mar. 15, 2025.

[17] C.-H. Chiu, T.-W. Huang, Z. Guo, and Y. Lin, "Pipeflow: An Efficient Task-Parallel Pipeline Programming Framework using Modern C++," arXiv, vol. 2202, no. 00717, pp. 1-13, Feb. 2022. doi: 10.48550/arXiv.2202.00717.

[18] S. Nazeema and S. G. Gundabatini, "Real-Time Text Detection and Translation Using OpenCV and Tesseract OCR: An Integrated Framework for Images and Videos," International Journal of Innovative Research in Technology, vol. 11, no. 6, pp. 2133-2139, Nov. 2024.

[19] J. Radas, B. Risse, and R. Vogl, "Building UniGPT: A Customizable On-Premise LLM-Solution for Universities," in Proceedings of EUNIS 2024 Annual Congress, Athens, Greece, 2025, pp. 108-116.

[20] M. Kalender, M. T. Eren, Z. Wu, O. Cirakman, S. Kutluk, G. Gultekin, and E. E. Korkmaz, "Videolization: Knowledge Graph Based Automated Video Generation from Web Content," Multimedia Tools and Applications, vol. 75, no. 24, pp. 4275-4294, Dec. 2016. doi: 10.1007/s11042-016-4275-4.

[21] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models," arXiv preprint arXiv:2109.10282, Sep. 2022.

[22] P. H. Jain, V. Kumar, J. Samuel, S. Singh, A. Mannepalli, and R. Anderson, "Artificially Intelligent Readers: An Adaptive Framework for Original Handwritten Numerical Digits Recognition with OCR Methods," Information, vol. 14, no. 6, pp. 305, May 2023. doi: 10.3390/info14060305.

[23] T.-Q. Lin, W.-P. Huang, H. Tang, and H.-Y. Lee, "Speech-FT: A Fine-tuning Strategy for Enhancing Speech Representation Models Without Compromising Generalization Ability," arXiv, vol. 2502, no. 12672, pp. 1-12, Feb. 2025. doi: 10.48550/arXiv.2502.12672.

Topologically Equivalent Binary Trees Regarding Binary Tree Roll

Adrijan Božinovski
Faculty of Informatics
American University of Europe
Skopje, Macedonia
adrijan.bozinovski@fon.edu.mk

***Abstract*—Binary trees T_1 and T_2 are topologically equivalent regarding operation P if $P(T_1) = T_2$, and P does not involve node insertion to or deletion from T_1. In other words, topological equivalence of T_1 and T_2 implies that the number of nodes in both trees is the same, but their structural arrangements differ. In this paper, the operation in question is the binary tree roll, and the functions for obtaining the Catalan Cipher Vector enumeration from and to a binary tree, as well as its ranking and unranking, are used. A clustering algorithm is presented, which is used to determine the clusters of topologically equivalent binary trees, and the number of those clusters for a given number of nodes. The numbers of those clusters are listed for binary trees with different numbers of nodes, and those results are discussed.**

Keywords—topologically equivalent binary trees, binary tree roll, Catalan Cipher Vector, rank, clustering algorithm

I. Introduction

A. State of the Art

A binary tree is a structure defined on a finite set of nodes that either contains no nodes or is composed of three disjoint sets of nodes, where the first set is a single node called the root node, the second set is itself a binary tree which is placed at the left subnode of the root and is called the left subtree, and the third set is also a binary tree and is placed at the right subnode of the root and is called the right subtree [1]. Binary trees are frequently used as auxiliary data structures when solving practical problems [2, 3, 4], but ongoing theoretical reseach into their properties is also being conducted [5, 6, 7].

The primary way to disambiguate the binary trees from one another is through knowledge of the strucutral arrangement of the nodes within the binary trees, i.e., what sub-nodes (left and/or right) are present in the root, then what sub-nodes are present in those sub-nodes etc. This structural arrangement of the nodes in the binary tree is commonly called the topology of the binary tree [7, 8, 9], and is a feature of every binary tree regardless of whether the nodes of the tree contain data fields or not. There are C_N distinct topologies for a binary tree with N nodes, where C_N is the N-th Catalan number [10].

The topology of a binary tree can be represented by an enumeration of it. The enumeration of a binary tree is a string of symbols, usually characters or integers, which unambiguously convey the tree's topology. There are various features based on which an enumeration of a binary tree can be obtained [11, 12, 13], but the two main groups of enumerations concern bit strings [12, 14, 15] and symbol sequences [16, 17, 18]. Some variants of the latter include enumerations using codewords [19, 20, 21], weights [22], distance [23], Catalan combinations [24] and Catalan Cipher Vectors, whereas codewords, Catalan combinations and Catalan Cipher Vectors are related, in a sense that it is possible to covert between any two of them [25]. The concept of ranking and unranking of the enumerations, i.e., assigning integer values to them, is also known [26, 27, 28] and will be used in this paper as well.

Some well known actions on binary trees are the binary tree traversals, determining the number of nodes in a binary tree, insertion and deletion of nodes in such a tree etc. Some of these actions don't change the internal structure of the binary tree – and hence its enumeration – and some of them do. For all of them the terms "function" and "operation" are occasionally used interchangeably, without disambiguation between the two (e.g., [1]). However, one way to disambiguate between them is to state that a function on a binary tree is an action that does not alter the tree's structure and hence its enumeration, whereas an operation on a binary tree is an action that changes the tree's structure and hence enumeration [29]. This disambiguation will also be adhered to in this paper, which means that herein the enumerations will be considered functions on binary trees.

If an operation P is performed on a binary tree T_1, the result will be a new tree T_2. This way, we can define that trees T_1 and T_2 are topologically equivalent, provided that no insertion or deletion of nodes has taken place. In other words, topological equivalence of T_1 and T_2 implies that the number of nodes in both trees is the same, but their topologies are different. The operation P can be of any type, e.g., [7, 30, 31], which allows for different binary trees to be topologically equivalent with one another regarding, i.e., when using, different operations.

B. Motivation and Setup

The focus of this paper are topologically equivalent binary trees regarding binary tree roll [29]. Specifically, the idea is to find out how many of all of the C_N binary trees with N nodes can be obtained from a single starting binary tree with a certain topology, by consecutively rolling it in a single direction. More importantly, the goal is to find out how many clusters of such topologically equivalent binary trees, regarding binary tree roll, are present, when incrementing the value of N.

Since the operation changes the topology of the binary tree, and since there is a 1-to-1 relationship between a binary tree's topology and its enumeration, representing binary trees using their enumerations is a valid way to address them, provided that the same enumeration is used for all binary trees that are considered. In this paper, the Catalan Cipher Vector enumeration will be used for that purpose, as well as its ranking and unranking procedures [25].

C.Structure of the Paper

The structure of the paper will be as follows. Chapter II will present the binary tree roll operation, as the operation regarding which topologically equivalent binary trees will be considered. Chapter III will present the Catalan Cipher Vector enumeration, i.e., how to convert a binary tree to a Catalan Cipher Vector based on its topology and vice versa, alongside the process of its ranking and unranking. Chapter IV shows the connections among the binary tree roll operations, functions concerning the Catalan Cipher Vector enumeration and the functions concerning the rank of that enumeration. Chapter V presents the clustering algorithm, for placing topologically equivalent binary trees regarding binary tree roll into clusters and counting how many clusters for binary trees with N nodes there are. Chapter VI presents the results of running that algorithm, alongside a discussion about some of the findings. The paper ends with Chapter VII, which is the Conclusion.

II.The Binary Tree Roll Operation

The binary tree roll is an operation in which a binary tree appears "rolled" for 90 degrees counterclockwise or clockwise, hence there are two variants of the binary tree roll, abbreviated $CCW()$ and $CW()$ respectively. For the binary tree roll to be defined, the preorder, inorder, and postorder traversals of the binary tree need to be available as functions $preorder()$, $inorder()$ and $postorder()$, respectively. In such a case, the counterclockwise roll is defined as

$$CCW(T_1)=T_2 \Leftrightarrow preorder(T_1)=inorder(T_2) \wedge inorder(T_1)=postorder(T_2) \quad (1)$$

and the clockwise roll is defined as

$$CW(T_1)=T_2 \Leftrightarrow inorder(T_1)=preorder(T_2) \wedge postorder(T_1)=inorder(T_2) \quad (2)$$

Some features of the binary tree roll include that both variants are inverses to one another, i.e., $CCW(T)=CW^{-1}(T)$ and $CW(T)=CCW^{-1}(T)$ and also that there is an algorithm to perform the binary tree roll without the need to know the traversals [29]. The original version of the binary tree roll algorithm is quadratic for time complexity [32] and space complexity [33], but an improved version of it makes it linear for both time complexity and space complexity [34].

III.Catalan Cipher Vectors, Their Ranking and Unranking

A.Definition

The Catalan Cipher Vector, or CCV, is an enumeration that corresponds to a binary tree's topology related to its level-order traversal [25]. If the binary tree has N nodes, the CCV is defined as the vector $[v_0 v_1 v_2 \dots v_{N-1}]$ with the features that $v_0=0$ and $v_{i-1}+1 \le v_i \le 2 \cdot i$ where $i=1,\dots,N-1$. There are C_N such vectors, which results in a 1-to-1 relation between CCVs with N elements and binary trees with N nodes.

B.Features

The values of the vector elements give information about which nodes they are connected to, specifically which are their parent nodes. Since the elements of the CCV give information regarding nodes as they appear in the level-order traversal of the binary tree, $v_0=0$ represents the root, whereas each subsequent CCV element represents each node of the tree as it appears in the level-order traversal of the binary tree. The value of each such element gives information about which is the index in the CCV of the parent node to the node at that index, and also about which sub-node – left or right – that particular node is to its parent node. More precisely, for each v_i where $i=1,\dots,N-1$, if v_i is odd then the node represented by v_i is the left sub-node of the node represented with the index $\frac{v_i-1}{2}$ in the CCV, and if v_i is even then the node represented by v_i is the right sub-node of the node represented with the index $\frac{v_i}{2}-1$ in the CCV. This enables the full reconstruction of the binary tree's topology, whereas the data stored in the elements may or may not be given (if they are, their sequence corresponds to the level-order traversal of the binary tree, and their data can be used to visually reconstruct the binary tree using the information in the nodes of the CCV). Since each node of the binary tree has its corresponding element in the CCV, the procedures for generating the CCV out of a given binary tree and also for recreating the binary tree from a given CCV are linear for both time and space complexity [25].

C.Ranking and Unranking

For each CCV with N elements it is possible to assign a unique integer value from 0 to C_N-1, called the rank of the CCV, through a process called ranking. Conversely, obtaining the CCV from a given rank is a process called unranking [25]. Both processes involve a special form of a Catalan Triangle, called the Bottom Ones Catalan Triangle or BOCT. It is obtained from the commonly considered default version of the Catalan Triangle [26] by omitting its longest diagonal, transposing it, and inverting it alongside the horizontal [28]. Table I shows the BOCT for $N=4$, and equation (3) shows how to calculate each element of it.

$$BOCT(r,c)=\frac{(r-c+2)\cdot(2\cdot N-r-c-1)!}{(N-r-1)!\cdot(N-c+1)!} \quad (3)$$

For each N, if the integer factorial values from $0!$ to $(2\cdot N-1)!$ are available to be fetched on demand, e.g., if they are pre-calculated and stored in an indexed data structure such as an array, the calculation of each element in the BOCT is done in constant time. This is the reason for the improvement of the space complexity of the ranking and unranking processes for the CCVs from quadratic [25] to linear [28], whereas the time complexity for both processes remains linear [25].

Table I. The Bottom Ones Catalan Triangle for $N=4$

r\c	0	1	2	3
0	14			
1	9	5		
2	4	3	2	
3	1	1	1	1

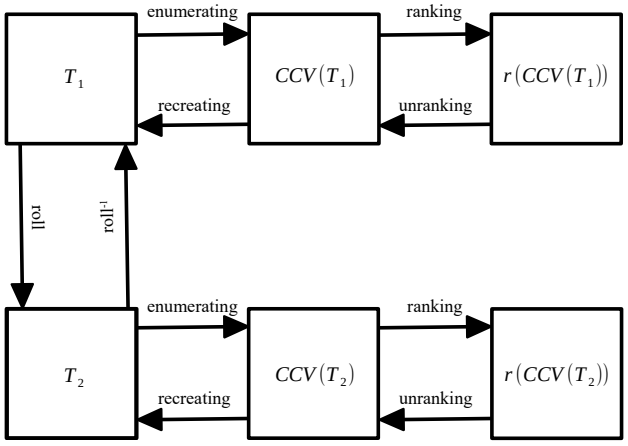


Fig. 1.The process of converting between topologically equivalent binary trees T_1 and T_2, as well as their corresponding enumerations and ranks.

IV.Converting between Topologically Equivalent Binary Trees Regarding Binary Tree Roll and Their Representations

Figure 1 shows the procedure for converting between topologically equivalent binary trees T_1 and T_2, as well as their representations, i.e., their CCV enumerations and ranks.

All of the actions represented as arrows in Figure 1 are linear for time and space complexity [25, 28, 34]. The "vertical" actions, i.e., the binary tree roll in both directions, are considered operations, since they change the topology of T_1 into T_2 and vice versa. Hence, the choice of the binary tree roll direction is arbitrary. The "horizontal" actions, on the other hand (i.e., the actions of enumerating, ranking, unranking and recreating), don't change the topology of the binary tree but rather represent it in different ways, and are therefore considered functions.

V.The Clustering Algorithm

A.A Case Study

Figure 2 shows the two clusters of topologically equivalent binary trees regarding binary tree roll with 3 nodes, i.e., for $N=3$. The trees that are in the same row belong to the same cluster. Viewing them from left to right they follow the counterclockwise roll, whereas viewed from right to left they follow the clockwise roll [29].

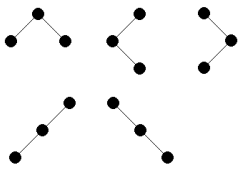


Fig. 2.Clusters of topologically equivalent binary trees regarding binary tree roll with 3 nodes.

Not all tree topologies are obtainable by consecutively rolling a binary tree with an arbirary starting topology. Therefore, it is necessary to devise a way to go through all of the $C_3=5$ binary trees and cluster all of them based on whether they are obtained as a result of the binary tree roll operation starting from a certain topology. Figure 3 shows the process of clustering all of the binary trees with 3 nodes, starting from the binary tree with rank 0, which is the full binary tree topology, according to the CCV enumeration and its ranking algorithm [25].

The process goes as follows. Starting from the binary tree with the CCV rank of 0, consecutively rolling it (in the counterclockwise direction, in this case) leads to the topology of the tree with rank 2, then 3, then back to 0. In this case, the starting rank of the cluster is reached again, and thus the current cluster gets finalized. Then, the first subsequent rank that is not in a cluster is reached – in this case the rank 1 – and a binary tree is created out of it. The aforementioned clustering procedure is repeated again, and this uses up all 5 of the binary trees with 3 nodes, which finishes both the second cluster and the algorithm.

B.The Pseudocode

Figure 4 shows the pseudocode for the clustering algorithm. The "keywords" of the pseudocode are written in normal letters, whereas the user created items, including the variables, the `BinaryTree` data type alongside its methods (`fromRank` – to obtain a new binary tree from a given rank; `r` – to return the rank for a given CCV [28]; `CCV` – to return the CCV for a given tree [25]; and `iCCW` – to return the binary tree obtained as a result of the improved, i.e., linear, CCW roll algorithm on a given binary tree [34]), procedure `initializeAllElements` to set all elements of an array to an initial value, and function `findNextFalse` to find the first subsequent false element's index in an array or `-1` if there is none, are written in italic.

The final line of the pseudocode outputs the size of queue `G`, which is used to store all of the clusters of topologically equivalent binary trees regarding binary tree roll. This information is used to determine how many such clusters there are for a given number `N` of nodes in the tree.

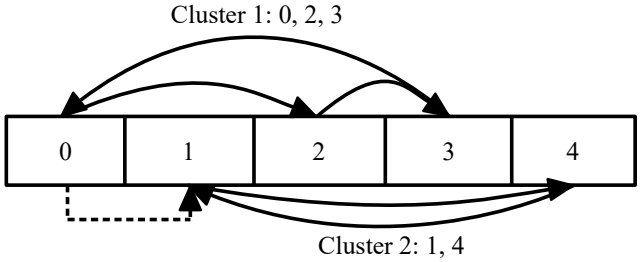


Fig. 3.Clustering the topologically equivalent binary trees with 3 nodes (shown in Figure 2) based on binary tree roll, where the trees are shown using their CCV ranks. The counterclockwise direction of the binary tree roll is used.

```
input(N)
C_N = (2*N)! / (N! * (N+1)!)
A = new array(bool, C_N)
initializeAllElements(A, false)
numTrue = 0
T = BinaryTree.fromRank(0)
g = new queue(int)
G = new queue(typeOf(g))
while(numTrue <= C_N)
   currentRank = BinaryTree.r(BinaryTree.CCV(T))
   if (!A[currentRank])
      A[currentRank] = true
      ++numTrue
      g.enqueue(currentRank)
      T = BinaryTree.iCCW(T)
   else
      G.enqueue(g)
      emptyQueue(g)
      currentRank = findNextFalse(A, currentRank)
      if(currentRank == -1)
          break
      T = BinaryTree.fromRank(currentRank)
print(G.numElements)
```

Fig. 4.The pseudocode of the algorithm for clustering of topologically equivalent binary trees.

VI.Results and Discussion

Table II shows the numbers of clusters of topologically equivalent binary trees regarding binary tree roll for given sizes of the binary trees. The results for binary trees with 19 and more nodes are unobtainable at the moment due to extensive memory requirements for storage as well as usage of large integer values. Further improvements will need to be made in order to enable such calculations.

Table II. Numbers of Clusters of Topologically Equivalent Binary Trees Regarding Binary Tree Roll

Number of nodes	Number of clusters
1	1
2	1
3	2
4	3
5	10
6	18
7	46
8	95
9	236
10	528
11	1288
12	2936
13	6984
14	16212
15	38528
16	90717
17	216648
18	516358

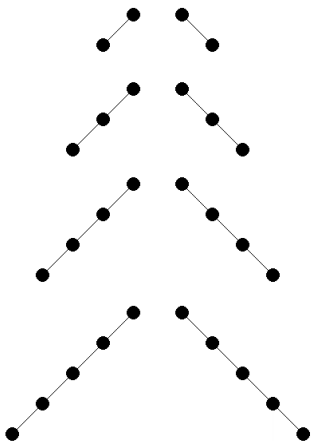


Fig. 5.Clusters of degenerate binary trees for $N=2\ldots5$.

Another notable result is the existence of a cluster of degenerate trees for every N , as shown in Figure 5, where the clusters of degenerate trees are shown for $N=2\ldots5$. This cluster is present for every N and always consists of only two elements, namely both the left- and right-degenerate tree with as many nodes. Knowing that the right-degenerate tree always has the rank of C_N-1 when using the CCV enumeration and ranking [25], the left-degenerate tree can be straightforwardly obtained by rolling the right-degenerate tree. Further research is needed to determine the importance of this cluster.

The Fundamental Theorem of Equivalence Relations states that an equivalence relation divides a set into equivalence classes, which is a label that might be tempting to assign to the clusters of topologically equivalent binary trees. However, in order for the binary tree roll to be considered an equivalence relation it would need to be reflexive, symmetrical and transitive [35]. While it might be argued that the binary tree roll is reflexive – given that its two directions of roll are inverse to one another – as well as symmetrical – for the same reason – it is not transitive (e.g., if $CCW(T_1)=T_2$ AND $CCW(T_2)=T_3$, then $CCW(T_1)=T_2\neq T_3$ and also $CW(T_3)=T_2\neq T_1$). Therefore, it would be incorrect to use the term “equivalence classes” to specify which binary trees are topologically equivalent to each other based on binary tree roll, and this is why the term “clusters” was used instead.

VII.Conclusion

This paper presents an approach towards clustering of topologically equivalent binary trees regarding binary tree roll, using the Catalan Cipher Vector enumeration and ranking procedures to disambiguate among the topologies. The concept of topologically equivalent binary trees regarding an operation has been presented, alongside an explanation of the binary tree roll operation, as well as the Catalan Cipher Vector enumeration and ranking processes. It is shown how they relate to the binary trees and an algorithm is presented, which has the goal of clustering topologically equivalent binary trees regarding binary tree roll. The results of that algorithm are presented for increasing numbers of nodes in a binary tree and a discussion is given for them.

REFERENCES

[1]T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 2009.

[2]A. Clifton, É. Czabarka, A. Dossou-Olory, K. Liu, S. Loeb, U. Okur, L. Székely, and K. Wicke. "Decks of rooted binary trees". European Journal of Combinatorics, vol. 124, 2023, 104076.

[3]A. Cinar, H. Iscan, and M. Kıran. "Tree-Seed Algorithm for Large-Scale Binary Optimization". KnE Social Sciences, vol. 3, pp. 48-64, 2018.

[4]B. Effantin. "Generation of Valid Labeled Binary Trees". *Computational Science and Its Applications — ICCSA 2003*. Lecture Notes in Computer Science, vol 2667, V. Kumar, M. L. Gavrilova, C. J. K. Tan, P. L'Ecuyer Eds. Springer, Berlin, Heidelberg, pp. 245-253, 2003.

[5]T. Wiehe. "Counting, grafting and evolving binary trees". *Probabilistic Structures in Evolution.* E. Baake, A. Wakolbinger Eds. pp. 427-450, 2020.

[6]I. Dramnesc, T. Jebelean, S. Stratulat. "Theory Exploration of Binary Trees". 13th IEEE International Symposium on Intelligent Systems and Informatics, SISY 2015, Subotica, pp.139-144, September 2015.

[7]H. Pajoohesh. "Topological and categorical properties of binary trees", Applied General Topology, vol. 9, no. 1, pp. 1–14, 2008.

[8]R. Grübel. "A note on limits of sequences of binary trees". Discrete Mathematics & Theoretical Computer Science, vol. 25, no.1. [Online] Available: https://doi.org/10.46298/dmtcs.10968, 2023.

[9]U. Mukherjee, and M. Gopi. "Finding Feature Similarities Between Geometric Trees". 22nd Pacific Conference on Computer Graphics and Animations. Seoul. pp. 73-78. [Online] Available: https://doi.org/10.2312/PGS.20141255, 2014

[10]The On-Line Encyclopedia of Integer Sequences, Sequence A000108, https://oeis.org/A000108, 2024.

[11]Y. Y. Kochetkov. "Enumeration of One Class of Plane Weighted Trees". Journal of Mathematical Sciences vol. 209, pp. 282–291. [Online] Available: https://doi.org/10.1007/s10958-015-2503-5, 2015.

[12]M. C. Er. "Enumerating Ordered Trees Lexicographically". The Computer Journal, vol. 28, no. 5, pp. 538-542, 1985.

[13]Z. Wang. "Additive Enumeration Problems Over Trees". Acta Mathematica Scientia, vol. 3 no. 4, pp. 445-453, 1983.

[14]A. Proskurowski. "On the generation for binary trees". Journal of the ACM, vol. 27, pp. 1-2, 1980.

[15]S. Zaks. "Lexicographic generation of ordered trees". Theoretical Computer Science, vol. 10, no. 1, pp. 63-82, 1980.

[16]G. D. Knott. "A numbering system for binary trees". Communications of the ACM, vol. 20, no. 2, pp. 113-115, 1977.

[17]F. Ruskey, and T. C. Hu. "Generating Binary Trees Lexicographically". SIAM Journal on Computing, vol. 6, no. 4, pp. 745-758, 1977.

[18]D. Rotem, an Y. L. Varol. "Generation of Binary Trees from Ballot Sequences". Journal of the ACM, vol. 25, no. 3, pp. 396-404, 1978.

[19]D. Zerling. "Generating binary trees using rotations". Journal of the ACM, vol. 32, no. 3, pp. 694-701, 1985.

[20]D. Roelants van Baronaigen. "A loopless algorithm for generating binary tree sequences". Information Processing Letters, vol. 39, no. 4, pp. 189-194, 1991.

[21]L. Xiang, C. Tang, and K. Ushijima. "Grammar-Oriented Enumeration of Binary Trees". The Computer Journal, vol. 40, no. 5, pp. 278-291, 1997.

[22]J. M. Pallo. "Enumerating, Ranking and Unranking Binary Trees". The Computer Journal, vol. 29, no. 2, pp. 171-175, 1986.

[23]E. Mäkinen. "Left distance binary tree representations". BIT Computer Science and Numerical Mathematics, vol. 27, no. 2, pp. 163-16, 1987.

[24]M. Črepinšek, and L. Mernik. "An Efficient Representation for Solving Catalan Number Related Problems". International Journal of Pure and Applied Mathematics, Vol. 56, No. 4, pp. 589-604, 2009.

[25]A. Božinovski, B. Stojčevska, and V. Pačovski. "Enumeration, Ranking and Generation of Binary Trees Based on Level-Order Traversal Using Catalan Cipher Vectors". Journal of Information Technology and Applications, vol. 3, no. 2, pp. 78-86, 2013.

[26]A. Proskurowski, and E. Laiman. "Fast enumeration, ranking, and unranking of binary trees". Proceedings of the Thirteenth Southeastern Conference on Combinatorics, Graph Theory and Computing, Boca Raton, FL, pp. 401-413, 1982.

[27]R. Wu, J. Chang, A. Chen, and M. Ko. "Ranking and Unranking of Non-regular Trees in Gray-Code Order". IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science, vol. 96-A, no. 6, pp. 1059-1065, 2013.

[28]A. Božinovski, and B. Stojčevska. "Full Linearization of Ranking and Unranking of Catalan Cipher Vectors Using Catalan Triangle Abstraction". Journal of Information Technology and Applications, vol. 14, No. 2, pp. 136-141, 2024.

[29]A. Božinovski, and N. Ackovska. "The Binary Tree Roll Operation: Definition, Explanation and Algorithm". International Journal of Computer Applications, vol. 46, no. 8, pp. 40-47, 2012.

[30]X. Guo, and M. Randic. "Trees with the same Topological Index JJ". SAR and QSAR in Environmental Research, vol. 10, no. 4, pp. 381–394, 1999.

[31]D. D. Sleator, R. E. Tarjan, and W. P. Thurston. "Rotation distance, triangulations, and hyperbolic geometry". Journal of the American Mathematical Society, Vol 1, No 3, pp. 647-681, 1988.

[32]A. Božinovski, G. Tanev, B. Stojčevska, V. Pačovski, and N. Ackovska. "Time Complexity Analysis of the Binary Tree Roll Algorithm". Journal of Information Technology and Applications, Vol. 6, No. 2, pp. 53-62, 2016.

[33]A. Božinovski, G. Tanev, B. Stojčevska, V. Pačovski, and N. Ackovska. "Space Complexity Analysis of the Binary Tree Roll Algorithm". Journal of Information Technology and Applications, Vol. 7, No. 1 pp. 9-19, 2017.

[34]G. Tanev, and A. Božinovski. "A linear time algorithm for rolling binary trees". Proceedings of the 17th IEEE International Conference on Smart Technologies IEEE EUROCON 2017, Ohrid, pp. 255-260, July 6-8 2017.

[35]D. S. Dummit and R. M. Foote. *Abstract Algebra Third Edition*. Wiley, 2004.

Implementing Interactive Learning For Educational Involvement Via Algorithmic Game Design

1st Marko Angelovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
angelovskim007@gmail.com

2nd Bojan Zhivkovikj
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
bojanzivkovik@gmail.com

3rd Andon Angov
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
angovandon@gmail.com

4th Viktor Denkovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
viktor.denkovski@uacs.edu.mk

5th Marija Stankova Medarovska
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
marija.s.medarovska@uacs.edu.mk

***Abstract*—The development of a game-based learning tool composed of a combination of web technologies (HTML and CSS) with a Flowgorithm as an interactive quiz game experience is presented in this study. QuizQuest presents a distinctive method of developing instruction-based games by combining interactive graphics with flowchart understanding. Well-known algorithmic solutions were implemented into the application, such as Bubble Sort for dynamic sorting questions by difficulty level and leaderboard ranking, a recursive algorithm for dynamic HTML content loading, and Linear Search for successful searching of players. The system covers many operations from user registration to tracking scores in the six-section architecture, ensuring an interesting and organised environment for learning. The study highlights how algorithmic thinking, data-driven interoperability, game-based learning tools and creativity may lead to improved educational approaches. Future implementation with AI-driven technologies would be considered to improve integration further and enable personalised learning experiences.**

Keywords—interoperability, Flowgorithm, VS Code, game-based learning, interactive teaching methods

I. Introduction

To improve student learning outcomes, these days it has become crucial to use engaging and interactive teaching methods. One of the most powerful approaches for educational purposes that raises motivation, information grasping, and engagement is game-based learning (GBL) [1]. It has been well proven and demonstrated that implementing educational games for the class improves the students' ability to think critically, solve problems, and apply knowledge [2]. To showcase an interactive educational involvement, this study presents QuizQuest, a game-based learning application that uses the interoperability between multiple well-known software and tools already integrated into academia for such purposes.

There were challenges to designing an application that could implement some usable functions from various software platforms that were created to be used for different operations, and try to connect their functionalities [3]. The game presented in this study was carefully designed by using Flowgorithm, an environment with visual programming capabilities which was mainly used for its user-friendly interface to create and program interactive visually represented algorithms using flowcharts [4], and Visual Studio Code (VS Code), an editor for source-code available for many different operating systems (macOS, Linux, and Windows) [5].

On the base level, this game presented an opportunity to showcase Flowgorithm's adaptability for programmers to apply flowchart logic to produce complex and dynamic applications that could support this game's architecture. Flowgorithm's ability to modify files stored in the computer helps with the creation of dynamic gameplay and graphics, which also provides users with an overview of the application's functionality [6].

On the other hand, VS Code presents a crucial tool to improve the capabilities of Flowgorithm's interaction with files with its built-in functionality for real-time file monitoring and modifications. For the game built in this paper, VS Code, supported with the Live Server extension, plays a significant part for the developers, and this challenge is to be taken as the responsible medium to introduce flexibility in merging with Flowgorithm, which enables real-time observations of code modifications.

Having to consider the above, building this game represents the complexity in the relationship between two different software platforms designed for different purposes, as it creates a rich learning experience that implements various well-known algorithms to successfully introduce the operations for the application. Adding to this, the fundamental difficulties shown with the game's integration, specifically regarding differences in performance between different devices, must be acknowledged [7]. Hence, the purpose of this study is to highlight how algorithmic thinking, game-based learning tools, creativity, and data-driven interoperability may improve educational approaches for students.

Game-based learning has been proven to be beneficial for raising motivation and positive outcomes by many studies, leading to being use for various educational purposes [8]. The six-section structure flow incorporated in QuizQuest was influenced by the famous quiz-style game show "Who Wants to Be a Millionaire?". By using the file operations within Flowgorithm on .txt and .html files, the application updates and processes the questions, answers, and leaderboards dynamically. This technique provides flexibility and interaction to the learning process that additionally introduces a variety of difficulty levels and enables the users to an immediate feedback.

This paper is structured into six sections, in which the first reviews the goals of the research. The second section discusses the related work done corresponding to this research. The third section goes over the materials and methods used for the creation of the game's core operations and functions. Continuing, the fourth section presents the results of the programming development, after which in the fifth section, the obtained results were discussed. Finally, section six concludes the paper.

II. Related Work

Many studies with implementations of game-based learning logic for educational purposes have shown that it can improve the engagement and performance of students. In [9], it was presented that video games encourage learning naturally by using context-based challenges and problem-solving tasks. Research shows how computational thinking can help students develop their problem-solving techniques [10]. Having said that, QuizQuest supports the combination of problem-solving algorithms with interactive learning and further demonstrates the ability to apply computational logic in education by incorporating real-time data updates of file operations. Similar to this, in [11] an investigation was done on the effectiveness of game-based learning approaches in computer science and shows that students who have used educational games had raised performances in knowledge retention and commitment in comparison to those who used conventional approaches. As studies explored the implementation of algorithmic designs and flow theory into games for education, in [12] it was focused on how motivation was raised by targeting the right balance between the level of skills and the difficulty of the game. QuizQuest directly applies this theory, having questions be constructed and chosen dynamically following the skill level of the user, which guarantees interest and replayability.

The algorithmic design of QuizQuest incorporates a key feature that is crucial for its ranking and dynamic question selection. To engage in a sufficient level of complexity for the students, this type of approach guarantees a progressive rise in difficulty [13]. According to research, when the difficulty level is increased, the learning results are greatly improved, corresponding to the user's performance in playing the game [14]. The organization of the process for selecting questions in QuizQuest supports these results, which dynamically modifies the difficulty of questions in accordance to the predetermined settings.

Similar to this research, in [15], a game-based learning project was also developed and introduced using the main software tools presented in this study and presented a well-designed multilayer platform with multiple mini-games like a maze, word search, towers of Hanoi, and tic tac toe. All this was presented by incorporating the algorithm-based logic of famous algorithms such as merge sort, linear and binary search and minimax recursive algorithm. This shows how introducing creativity to problem-solving techniques can stimulate students to develop and create something out of the ordinary and use it for learning purposes.

III. Materials And Methods

The main motivation behind this project was to improve the learning skills by trying to find a way to connect Flowgorithm, which was used as a base to develop the logic of the game by implementing popular sorting, searching, and recursive algorithms with VS Code, to not only make a fun game but also a visually appealing one.

As such, different types of algorithms were used and implemented in Flowgorithm to make an interesting quiz game consisting of questions with increasing difficulty. To load the question, Flowgorithm's file reading/writing capabilities were used. By exploiting the ability to read and write files, a system where Flowgorithm would write to specific HTML files to load the graphical user interface (GUI) was created. This HTML was then loaded on a local web server by the Live Server extension.

A. Used Technologies

The development of QuizQuest required the following tools:

- Flowgorithm: Algorithmic design and simulation tool [4];
- Visual Studio Code: HTML, CSS, and JavaScript-based UI design [5];
- Live Server Extension: Real-time web updates for interactive play [16];
- Data Files: .txt files for question storage and leaderboard tracking.

The entire game's structural software architecture design used for QuizQuest is presented in Fig. 1.

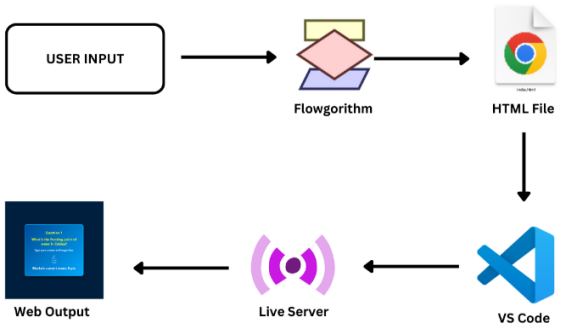


Fig. 1. Diagram of the QuizQuest software architecture

B. The Approach

The game follows structured algorithmic principles, with a hierarchical function-based design. The system uses bubble sort for question difficulty sorting and file reading/writing for dynamic updates. For this, the following was considered for:

- Data Processing and Visuals in which: Questions are stored in QuestionsPlus4Answers.txt, with each entry structured into seven lines (question, four answers, correct answer, and difficulty level). The questions are sorted using Bubble Sort, ensuring an increasing difficulty order. During gameplay, questions are selected from the dataset using randomized indexing. Flowgorithm manages logic and computations. HTML and JavaScript dynamically update the visuals, and finally, Live Server enables real-time UI updates.
- Game Functionality in which: Users interact with Flowgorithm to input answers, the system dynamically updates the game screen using readhtml() functions, and after completing the game, scores and names are saved and displayed.

- Leaderboard & Search Functionality in which: Bubble Sort arranges the top 10 scores, a linear search algorithm enables specific user lookups, and data is continuously written to scores.txt and names.txt for persistent storage.

IV. Results

A working quiz game with an interactive and fun user interface was developed. The input of the program was the Flowgorithm console. The output and visual aspect of the game were loaded in a browser window, which was constantly being updated based on the user's input, as shown in Fig. 2.

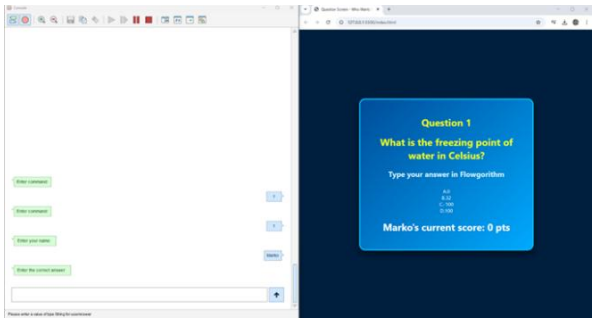


Fig. 2. QuizQuest game interface

As such, the Game Flow was organised as follows:

- A Main Menu, including Start Game, View Leaderboard, and Search User;
- The Gameplay, including Seven Questions, Randomized Selection, and Answer Validation; and
- Game Over, including Score Display, and Leaderboard Update.

The user provides input through the Flowgorithm console. The visual interface on the right is an index.html file (Fig. 2), which is open in VS Code using the Live Server extension to receive real-time updates. When the user provides input to the Flowgorithm console, the Flowgorithm file, which handles the logic, goes through and reads the needed pre-made .html file and stores it in a variable. It then erases the contents of the index.html file and replaces it with the contents of said variable. When VS Code detects changes to the index.html file, the Live Server extension refreshes the page and displays the updated information.

At the start, the Main Section of the program is loaded. It displays the game's title and allows the user to continue forward. After the appropriate input is given, the game's Main Menu is loaded. Here, the user can choose from one of three options: start the game, view the leaderboard, or search for a user score as shown in Fig. 3. Depending on the user's input, the appropriate screen is loaded.



Fig. 3. Main Menu page displayed through Live Server Extension

If the user chooses to load the leaderboard the top ten highest scores stored in the text file are loaded and displayed on the screen as presented in Fig. 4.



Fig. 4. Presented Leaderboard from QuizQuest

If the user decides to start a new game, the quiz section is loaded. Here, questions are randomly chosen from a preloaded text file, and after every correct answer, the difficulty of said question is increased. At the end of the game, the Game Over screen is shown, which displays the final score and updates the user's data by placing their name and score in a text file as shown in Fig. 5.

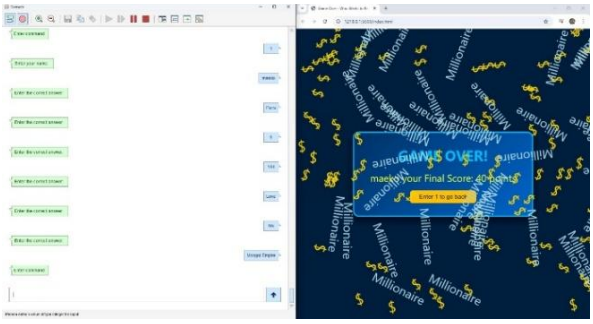


Fig. 5. Finalising the game with a game over screen with the score

The final option given to the user in the Main Menu is to load the Search User screen, where the user can search for a specific player's score via their name and have it displayed on the screen as shown in Fig. 6.

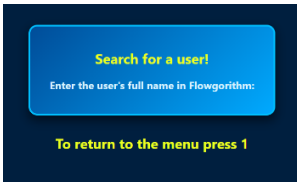


Fig. 6. A feature to search for a specific user's score

V. Discussion

This study demonstrates how a well-designed game-based learning tool may raise students' engagement and retention of information. By combining interactive gameplay and computational design, the provided methods allowed for the successful creation of a dynamic environment for learning. Flowgorithm and VS Code enabled the processing of structured data, and the questions were saved in organised structured files that were sorted and presented dynamically. For progressing the learning experience and sorting the

questions according to the difficulty level, the Bubble Sort algorithm was crucial [9, 17]. The leaderboard feature further contributed to raising the learner's motivation [8, 11].

The question selection and data processing presented in QuizQuest, are two of its many highlighted benefits. Furthermore, the system guarantees an interesting and dynamic user experience with the implementation of web-based visuals and real-time operations on files. Although having all this in mind, there are certain drawbacks and limitations to this solution, such as the requirement of implementing more advanced algorithms and flexible ways of selecting questions to improve user accessibility [13, 14]. Furthermore, although the current UI/UX operate successfully, user engagement and implementation could be increased with additional improvements, such as an AI-driven question selection process and an upgraded UI/UX could help increase the learning experience to the next level [18]. Implementing machine learning algorithms would improve question selection by examining the user performance trends and dynamically changing the difficulty level. By using strategies such as collaborative filtering or reinforcement learning, the system could tailor the quiz play sessions to the user's skill level and the speed of learning. Comparing this method to the already implemented static difficulty scaling, it would present a more flexible and interesting experience, which would guarantee the learners are pushed without overburdening them.

VI. Conclusion

In this paper, a non-conventional merger of two software platforms designed for totally different functionalities was introduced. A game-based learning platform for computer science education was demonstrated in this research. The developed system offers a dynamic and elevated learning experience by combining a structured selection of questions, algorithmic processing, and an interactive game. Future research will concentrate on replacing Flowgorithm with a more complex-driven programming language for more advanced algorithms to be implemented, like Python, UI/UX improvements, and AI-driven adaptation in learning. Integrating machine learning could overcome limitations such as customization restrictions and static selection of questions. To improve the engagement of students, implementations such as badges for various achievements, monitoring of players' progress, and interactive animations could be implemented. Furthermore, accessibility features like customizing font styles and sizes, screen resolutions, and high contrast mode would provide a more personalized experience for various users. Considering all the above, QuizQuest paves the groundwork for the development of the next interactive learning and teaching materials. Considering all of the points presented in this research paper, the software platform Flowgorithm should not be used for the production of large and serious applications. This tool is best used for learning purposes and idea sharing, as it is not optimised for complex operations. It is a great beginner tool for computer science students to see and better understand program flow and experiment with program interaction. Because of Flowgorithm's insecure and slow nature, implementing complex functionalities in Flowgorithm should be avoided.

References

[1] Plass, J. L., Homer, B. D., and Kinzer, C. K. "Foundations of Game-Based Learning." Educational Psychologist, vol. 50, no. 4, 2015, pp. 258–83.

[2] Chen, L., Chen, P., and Lin, Z. "Artificial Intelligence in Education: A Review." *IEEE Access*, vol. 8, no. 8, Apr. 2020, pp. 75264–78.

[3] Ludwig, T., Wismueller, R., Sunderam, V., and Bode, A. "OMIS --On-line Monitoring Interface Specification (Version 2.0)", LRR-TUM Research Report Series, [online] 2, pp.173–180, 2004, Available at: https://www.cyfronet.pl/cgw05/contr-papers.pdf.

[4] "Flowgorithm - Flowchart Programming Language." *Www.flowgorithm.org*, www.flowgorithm.org/.

[5] Microsoft. "Visual Studio Code." *Visualstudio.com*, Microsoft, 2024, code.visualstudio.com/.

[6] Kourouma, M. K. "Capabilities and Features of Raptor, Visual Logic, and Flowgorithm for Program Logic and Design.", 2016.

[7] Awwal, N., Scoular, C., and Griffin, P. "Implementation issues with games-based assessment", EDULEARN16 Proceedings, [online] pp.905–910, 2016.

[8] Alotaibi, M. S. "Game-Based Learning in Early Childhood Education: A Systematic Review and Meta-Analysis." *Frontiers in Psychology*, vol. 15, no. 1, Frontiers Media, Apr. 2024.

[9] Chiotaki, D., Poulopoulos, V., and Karpouzis, K. *Adaptive Game-Based Learning in Education: A Systematic Review*. June 2023.

[10] Lester, D., Skulmoski, G. J., Fisher, D. P., Mehrotra, V., Lim, I., Lang, A., and Keogh, J. W. L. "Drivers and Barriers to the Utilisation of Gamification and Game-Based Learning in Universities: A Systematic Review of Educators' Perspectives." *British Journal of Educational Technology*, Feb. 2023.

[11] Papastergiou, M. "Digital Game-Based Learning in High School Computer Science Education: Impact on Educational Effectiveness and Student Motivation." *Computers & Education*, vol. 52, no. 1, Jan. 2009, pp. 1–12.

[12] Videnovik, M., Vold, T., Vibeke Kiønig, L., Madevska Bogdanova, A., and Trajkovik, V. "Game-Based Learning in Computer Science Education: A Scoping Literature Review." *International Journal of STEM Education*, vol. 10, no. 1, Springer Science+Business Media, Sept. 2023.

[13] Shute, V. J., Ventura, M., and Kim, Y. J. "Assessment and Learning of Qualitative Physics in Newton's Playground." *The Journal of Educational Research*, vol. 106, no. 6, Nov. 2013, pp. 423–30. Accessed 14 Jan. 2020.

[14] Van der Kleij, F. M., Feskens, R. C. W., and Eggen, T. J. H. M. "Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes." *Review of Educational Research*, vol. 85, no. 4, Dec. 2015, pp. 475–511.

[15] Kostadinov, N., Vuchkovikj, A., Denkovski, V., and Stankova Medarovska, M. "'Cryptic Chambers' Game - Interactivity and Interoperability Using Flowgorithm and Visual Studio Code." *21th International Conference on Informatics and Information Technologies (CIIT 2024)*, Apr. 2024.

[16] "Live Server - Visual Studio Marketplace." *Marketplace.visualstudio.com*, marketplace.visualstudio.com/items?itemName=ritwickdey.LiveServer.

[17] Resnick, M., and Robinson, K. "Lifelong Kindergarten: Cultivating Creativity through Projects, Passion, Peers, and Play". The Mit Press, 2017.

[18] Mustafa, M. Y., Tlili, A., Lampropoulos, G., Huang, R., Petar Jandrić, Zhao, J., Salha, S., Xu, L., Panda, S., None Kinshuk, Sonsoles López-Pernas, and Saqr, M. "A Systematic Review of Literature Reviews on Artificial Intelligence in Education (AIED): A Roadmap to a Future Research Agenda." *Smart Learning Environments*, vol. 11, no. 1, Springer Nature, Dec. 2024.

A Custom Mocking Server for API Simulation: Design, and Implementation

1st Rilind Kaba
Faculty of Informatics
Mother Teresa University
Skopje, North Macedonia
rilind.kaba@students.unt.edu.mk

2nd Fisnik Doko
Faculty of Informatics
Mother Teresa University
Skopje, North Macedonia
fisnik.doko@unt.edu.mk

3rd Martin Djonov
Faculty of Computer Science
University of Skopje
Skopje, North Macedonia
m.djonov@utms.edu.mk

4th Stojan Kitanov
Faculty of Informatics
Mother Teresa University
Skopje, North Macedonia
stojan.kitanov@unt.edu.mk

***Abstract*— Mocking servers are essential tools for simulating API behavior, enabling developers to test applications without relying on external services that are often unavailable or incomplete. These servers also allow developers to build applications even when other parts of the system are not yet ready. This paper introduces a custom mocking server designed to offer dynamic request handling, customizable response templates, and support for diverse data formats, the server provides enhanced flexibility and adaptability. Additionally, we present the server architecture, highlighting its ease of integration into existing testing workflows. In the end we will discuss potential improvements and broader applications, emphasizing the role of this solution in modern software development, particularly for API testing, as well as API and UI development. This research introduces a versatile tool that empowers development teams with greater control over their testing environments .**

Keywords—Mocking Server, RESTful APIs, HTTP Methods, Response Codes, URLs, API, API Simulation, .NET, Docker

I. Introduction

API or application programming interface, can be defined as a set of protocols that help data transmission and communication between different software components. APIs are being used from developers in order to create applications that are able to meet user needs by filling in the gaps between different and small pieces of code [1]. They define functionalities independent of implementations, allowing for flexibility and easier program development by offering reusable building blocks [2]. This modular approach to development helps to streamline software creation, as developers can focus on leveraging existing API components rather than reinventing the wheel each time a new feature is required. Furthermore, APIs play a critical role in enabling interoperability between diverse systems, allowing them to exchange information regardless of their underlying technology stacks.

Developers and testers use API testing tools, which are important because they help them in verifying the functionality, performance, security, data accuracy and response formats of APIs. The testing and simulation of APIs are critical stages in the software development lifecycle, especially for applications that depend on external services. Api testing is crucial for ensuring functionality, reliability, and security of software applications, particularly in distributed environments [3]. However, it presents several challenges, including sequencing API calls, handling unpredictable responses, and managing parameter dependencies [3]. These challenges require robust testing frameworks and careful planning to simulate real-world scenarios effectively. Additionally, thorough testing helps identify edge cases and prevent issues that may arise under high load or unusual conditions..

Mock servers play a crucial role in API testing by simulating real server behaviour, enabling developers to test applications without relying on actual backend services [4], [5]. They can simulate a variety of response scenarios, such as error codes or latency, helping developers prepare for a range of possible outcomes without having to rely on live production systems. This helps accelerate the development cycle, as testing can continue even when the actual API is unavailable or not yet fully developed.

Moreover, mock servers are valuable when dealing with incomplete or unstable third-party APIs, as they provide a stable, predictable environment for developers to work in while the external services are still being finalized or improved.

II. Motivation

The motivation for this research stems from practical experience working with microservices and cross functional teams.

In many instances, front-end development cannot proceed without the backend being available. This challenge can be mitigated by mocking specific backend endpoints, which allows the front-end team to continue their work in parallel with the backend development. This approach minimizes delays and reduces dependencies between teams, ultimately enabling a more efficient development process.

Additionally, there have been situations where it was not possible to stop other services during testing, making it difficult to test certain scenarios thoroughly. While third-party services exist to provide endpoint mocking, their use often involves external dependencies and the risk of exposing sensitive data.

The proposed solution, however, can be deployed directly within the project, ensuring that no sensitive information is compromised while facilitating efficient parallel development and testing.

III. The Objectives

The primary objectives of this project are:

A. Enable Parallel Development

To develop a mock service that simulates specific backend endpoints, allowing front-end teams to proceed with development even when the backend is unavailable.

This will minimize downtime and dependencies between teams, ensuring that the front-end can continue its work without waiting for backend readiness.

B. Improve Testing Efficiency

Microservices architecture presents unique challenges for API testing, requiring advanced strategies to ensure system reliability. Key difficulties include managing inter-service dependencies, handling asynchronouse communications, and mantaining data consistency across distributed nodes [6]. Creating a solution that allows for more comprehensive testing of microservices by enabling the simulation of services that may not be immediately available or easily stoppable.

This will improve the overall test coverage and reduce bottlenecks during the testing phase.

C. Enable data securit, privacy and compliance

To develop a mock service solution that can be deployed within the project environment, avoiding the exposure of sensitive data associated with third-party services.

This mock service will act as a controlled environment for simulating interactions with external RESTful APIs, ensuring that developers can test functionality without risking exposure of sensitive data. By using mock services, developers can simulate responses, error scenarios, and other aspects of third-party services while keeping the real data secure.

Information security compliance is a critical concern for organizations, particularly in meeting SOC 2 Type II requirements [7], [8]. Organizations must protect sensitive data, including personally identifiable information (PII) and protected health information (PHI) to comply with regulations like HIPAA and GDPR [9]. The Payment Card Industry Data Security (PCI DSS) is a crucial benchmark for protecting cardholder data in IT systems [10]. Implementing a mock service solution enables organizations to test applications without exposing real sensitive data, reducing legal and compliance risks.

Moreover, by isolating development from real production environments, this approach helps reduce the risk of data breaches during the development cycle. This solution will help maintain security standards while enabling efficient development.

D. Seamless Integration

To design the mock service in a way that integrates seamlessly into existing development workflows, ensuring minimal setup and configuration effort, while providing the flexibility to easily switch from mocked endpoints to live services when available.

E. Enhance Collaboration

To foster better collaboration between front-end and back-end teams by reducing the friction caused by interteam dependencies. By enabling front-end development to continue in parallel with backend work, the overall development cycle can be accelerated.

IV. Mock Server

Mock server is a tool used in development, integration and testing to guarantee efficient software workflows. It is widely used because it allows developers to simulate the behavior of real APIs and helps them to perform testing and debugging. This enables faster iteration, as developers can continue working on their application even if the real API is not available or not fully developed.

Simulation servers based on API specifications can bridge the gap between frontend and backend development, allowing for testing when the backend is not fully implemented[11]. This enables frontend teams to continue development even when backend APIs are incomplete. A mock server is essentially a simulated server environment created to mimic the actions of a real server, and while a real server can introduce complexity during development and testing because of external dependencies, developers use mock servers to establish an environment where they can work without waiting for external services [12]. By simulating different API responses—such as success, errors, timeouts, or specific HTTP status codes—mock servers provide a controlled environment for testing edge cases and failure scenarios, which might be hard to replicate with live APIs. Mock servers reduce dependency on backend teams, allowing frontend developers to simulate API responses based on predefined data or API specifications. The ability to simulate various response times and error conditions ensures that applications are robust and resilient before being deployed to production, helping to identify issues early and improving overall testing reliability.

V. Existing Mocking Tools

A. Postman

Postman is one of the most widely used tools for API testing [13], and it includes a mock server feature. This feature allows developers to simulate API endpoints by defining responses for specific requests. It simplifies the process for developers by providing a user-friendly interface, making it an ideal choice for smaller projects or less complex use cases.

A Mock Server in Postman is a feature that allows you to simulate API responses without needing the actual backend to be functional. This is particularly useful in development and testing when the backend is not yet developed, or when simulating edge cases and error conditions.

Here’s how the Mock Server in Postman works:

Creating Mock Servers: Postman allows you to create mock servers directly from collections or individual requests, you can define the mock response (status code, headers, body data) for each request in the collection [14].

Simulating Responses: When a request is sent to the mock server, Postman returns predefined responses based on the specifications in the mock server. This helps frontend developers or other stakeholders to continue their work even when the actual API is not yet available[14].

Using Mock Server URLs: Once the mock server is set up, Postman provides a mock URL that can be used to send requests. This allows users to simulate interactions with the API as if it were live.

Customizing Mock Responses: You can customize the mock server responses for different conditions, such as success or error scenarios, to thoroughly test your API without needing the actual backend.

Testing Edge Cases: Mock servers are perfect for testing scenarios where you need to simulate server errors, slow response times, or other conditions that might be difficult to reproduce with a live server.

B. *WireMock*

WireMock is a more feature-rich solution designed. WireMock is a tool for API mocking and service virtualization that allows developers and testers to simulate API endpoints and services during development and testing.

It's often used when the actual APIs or external services are unavailable, incomplete, or difficult to work with. WireMock can simulate the behavior of real APIs, allowing testing and development to continue smoothly [15].

WireMock Cloud is a managed version of the WireMock tool that is hosted on the cloud. It simplifies the process of mocking APIs by providing an easy-to-use web interface, making it more accessible for teams without worrying about server management or local setup [16].

In WireMock Cloud, mock data is typically stored in the cloud (on WireMock's cloud infrastructure). This allows users to create, manage, and share mock APIs across teams [17].

VI. Our Custom Solution

Our custom mocking server, developed using .NET Core and Docker, is specifically designed to be integrated directly into your project, prioritizing data privacy. Unlike other third-party tools, which operate externally, our solution ensures that all data remains within your local, on-premises, or cloud environment—wherever you choose to deploy it. The key advantage is that the data stays within your architecture, not in the third party, minimizing any potential risks associated with data exposure.

The solution allows for easy creation and management of mock routes and responses through a simple web interface, which can be seamlessly integrated into larger continuous integration/continuous deployment (CI/CD) pipelines. The dynamic response generation and artificial delay simulation features are essential for accurately testing real-world conditions, such as variable response times or erratic network behavior.

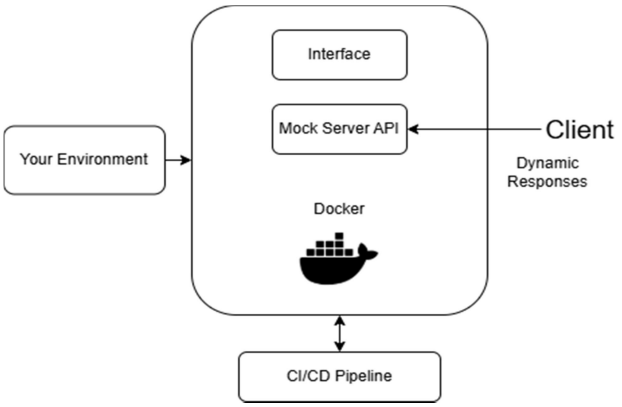


Fig. 1. Diagram of CustomMock Solution

Moreover, the use of Docker for deployment ensures that the mock server runs consistently across different environments, whether it's a local development machine, staging server, or a production-like environment. This consistency reduces the complexity and potential errors when transitioning between environments, enabling teams to focus on testing and development without worrying about discrepancies between environments. Docker's ability to isolate the mock server in a container also means that it can easily be scaled, replicated, or versioned, providing a high degree of flexibility when managing different instances or versions of the server.

Furthermore, this solution supports the HTTP protocol, making it ideal for mocking RESTful APIs, web services, or any system that communicates over HTTP. While the focus is on HTTP, the server can also support HTTPS, ensuring secure communication if needed. The server can provide a wide array of response simulations, including different HTTP status codes and responses, allowing for comprehensive testing and validation of client-server interactions in a controlled environment.

Docker has emerged as a leading containerization technology, offering significant advantages in efficiency and scalability for managing and deploying applications [17].

VII. System Architecture

The architecture of the mock server is designed to be modular, scalable, and easy to manage. It consists of several key components: the User Interface (UI), the Mock Server RESTful API, the Configuration Database, and the Docker Container

A. *User Interface*

User Interface is a web-based interface that allows users to manage and configure the mock server with ease. User can create APIs and for these APIs users can create, modify, and delete mock routes on form-based approach.

B. *Mock Server API*

The core component of the system is the Mock Server API, which is built using ASP.NET Core.

This RESTful API handles incoming HTTP requests, matches them to predefined routes, and serves the appropriate responses. The mock server allows for easy simulation of various HTTP status codes (e.g., 200 OK, 404

Not Found, 500 Internal Server Error) and predefined response bodies for different routes.

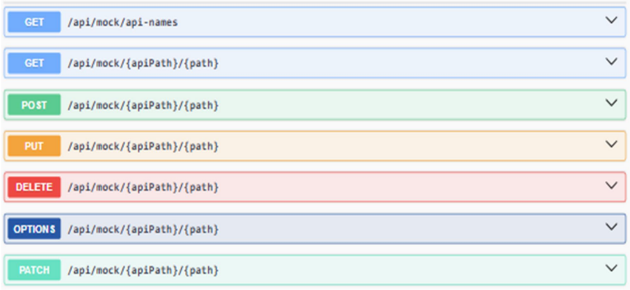


Fig. 2. Avaliable endpoints in Swagger

API First Development, which emphasizes designing API before implementation, is gaining popularity due to trends like microservices and mobile applications [18]. This approach aligns with modern software development practices, where RESTful APIs serve as the primary communication layer between different services. By defining the API first, teams can focus on designing a consistent and well-documented interface, ensuring smooth integration between components. The RESTful API integrates with Swagger/OpenAPI, this allows users to interact directly with the mock server via a Swagger-based UI. The OpenApi/Swagger specifications has become the standard for describing web APIs, benefiting various software development activities particularly improving documentation [18]. Through this interface, user can send HTTP requests, and receive the predefined mock responses. This integration streamlines the testing process by enabling users to quickly check the expected behavior of the mock server.

C. Database

SQL (Structured Query Language) remains a cornerstone of relational databases management systems offering a standardized approach to data manipulation and retrieval [19]. The database serves as a central repository for storing all API route definitions, including endpoint paths, supported HTTP methods (GET, POST, PUT, DELETE,OPTIONS, PATCH), and predefined response bodies. By managing route configurations within the database, the system enables dynamic updates without requiring code changes, improving flexibility and scalability.

D. Docker Container

Docker has become a popular containerization solution, enabling efficient application packaging and deployment [20]. Using Docker to containerize the entire mock server provides several benefits. Docker ensures that the application runs consistently across different environments by isolating the application from host machine configurations. This makes it easy to deploy the mock server to any environment (development, staging, production) without worrying about underlying OS or dependencies. Docker also simplifies integration with other tools in the development pipeline.

E. CI/CD Pipeline

CI/CD pipelines are crucial for enhancing efficiency in IT projects,reducing deployment failures, improving error detection, and facilitating faster release cycles [21]. Implementing CI/CD pipelines can lead to increased productivity, shorter time-to-market, fewer manual errors, and better collaboration between development and operation teams [22].

The server supports seamless integration into CI/CD pipelines, streamlining the entire deployment process. By automating the build, testing, and deployment stages, the server ensures that new updates are consistently delivered with minimal manual intervention.

VIII. Functional Overview

A. User Authentication and API Dashboard

Upon successful login, users are granted access to a central dashboard where they can manage all of their APIs.

After logging in, users are presented with a list of all available APIs, along with the option to view detailed information about each API, such as the associated routes and their respective configurations.

B. API Creation and Management

Users can create new APIs by specifying the APIs name and base path (e.g., /v1/users). This provides a structured method for managing APIs within the system.

Once created, APIs can be further customized by adding various routes, each defined by an HTTP method (GET, POST, PUT, DELETE,OPTIONS, PATCH) and a specific path (e.g., /users/list). This flexibility allows users to create any necessary endpoints for their mock API service.

C. Mock Route Configuration

The core functionality of the system is the ability to define mock routes for each API. A mock route is defined by the following parameters:

HTTP Method: The type of HTTP request (e.g., GET, POST, PUT, DELETE,OPTIONS, PATCH).
Route Path: The path for the specific endpoint (e.g., /users/list).
Request Body (for all HTTP requests): The structure and content expected in the request payload.
Response Body: The mock data that should be returned in the response, typically in JSON format.
HTTP Response Code: The HTTP status code that indicates the result of the request (e.g., 200 OK, 404 Not Found).
These mock routes can be updated or deleted as needed, providing users with full control over their simulated API interactions.

D. Swagger Integration for Dynamic API Documentation

A Swagger UI interface is integrated into the system, accessible via the /swagger route. This UI serves as an interactive API documentation tool, dynamically displaying all API endpoints, including the paths and HTTP methods. Users can interact with the mock routes directly from the Swagger interface, sending simulated requests to the mock APIs and receiving responses based on the configured mock data.

In addition to displaying individual API routes, Swagger exposes a special endpoint, such as /api/mock/api-names, which returns a comprehensive overview of all the APIs and their associated routes. This endpoint provides users with detailed metadata for each API, including:

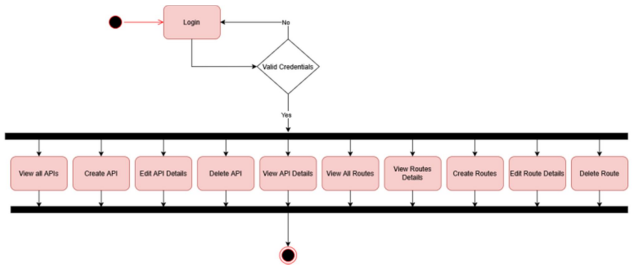


Fig. 3. How the CustomMock works

API Name: The human-readable name of the API.
Base Path: The base path for the API (e.g., /v1/users).

IX. Integration

In many cases, simulating the behavior of external services can streamline development, improve testing flexibility, and provide a controlled environment for testing various edge cases and failure scenarios.

However, instead of directly calling Microservice 2, Microservice 1 can be configured to send its request to the custom mock server. This mock server simulates the behavior of Microservice 2, providing a predefined response that is expected from the real service. As a result, Microservice 1 can receive the desired result—whether it's data, a success message, or a specific HTTP status code—without actually interacting with Microservice 2.

This setup is particularly useful during development and testing phases, where developers can simulate different behaviors, test various scenarios, and reduce dependencies on the real services.

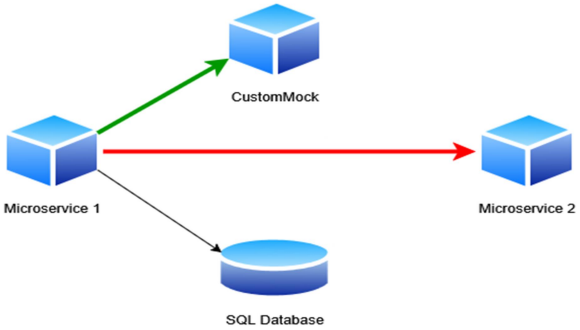


Fig. 4. Architecure of Usage the CustomMock in Project

X. Performance Results

Scala is a high-level, JVM-based programming language that combines object-oriented and functional programming [23], while Gatling is a load-testing tool written in Scala for simulating high-performace user traffic on web applications [24].
The simulation begins with a sharp surge of 300 users within a single second, designed to assess the system's ability to handle sudden, high-demand spikes. Following a 5-second pause for stabilization, the load increases gradually by 1,000 users over the course of 5 seconds, providing insight into the system's capability to manage sustained, incremental growth in activity. As shown in Fig. 5 throughout the entire test, all requests were processed

Requests	Executions				
	Total	OK	KO	% KO	Cnt/s
All Requests	7800	7800	0	0%	150
GET Request	1300	1300	0	0%	25
POST Request	1300	1300	0	0%	25
PUT Request	1300	1300	0	0%	25
PATCH Request	1300	1300	0	0%	25
DELETE Request	1300	1300	0	0%	25
OPTIONS Request	1300	1300	0	0%	25

Fig. 5. Performance tests results

successfully, demonstrating robust performance under both sudden and steady loads.

XI. Future Directions and Improvements

While the current implementation of the custom mock server is highly effective, there are several opportunities for improvement.

One potential enhancement is the integration of artificial intelligence to automate the generation of mock responses. Recent research expolores the use of generative AI for test data generation across various domains. [25] Large Language Models (LLMs) capability to generate realistic test data at different levels of integrability, from raw data to programs using fake libraries. This could significantly streamline the testing process by creating dynamic, context-aware mock responses that are tailored to the specific needs of the application.

Additionally, the use of AI could reduce the manual effort involved in creating and updating mock responses, making the mock server more adaptable and responsive to changes in requirements. By analyzing responses provided models, the system could predict and generate realistic responses, further enhancing the accuracy of the simulation. This would create a more dynamic and intelligent mock server, which could adjust responses based on real-world trends. Moreover, the AI could simulate a broader range of real-world scenarios, including edge cases and complex interactions, which might be difficult to manually define.

Another promising improvement is the support for GraphQL in addition to traditional REST APIs. GraphQL is emerging as an important alternative to Rest API for modern web applications. It offers greater flexibility and efficiency by allowing clients to request specific data in a single query, reducing over-fetching and under-fetching issues common with REST [26], [27]. As the adoption of GraphQL continues to grow, supporting it in the mock server would allow developers to work with both traditional REST APIs and more modern, flexible query languages. This would make the mock server even more versatile, accommodating a wider range of use cases and providing a more comprehensive solution for developers. Supporting GraphQL in the mock server would add an extra layer of complexity and value, particularly for developers working with modern API architectures.

Finally, incorporating advanced analytics and monitoring capabilities into the mock server could provide valuable insights into usage patterns. In the context of software product line architectures, service utilization metrics can be employed to assess and improve architectural quality [28]. These analytics could be used to track not only performance metrics like response times but also usage trends and

potential bottlenecks in the mock servers workflow. By visualizing data on how the mock server is being used, development teams could identify potential improvements in both the mock servers functionality and the overall development process. By tracking metrics such as request frequency, response times, and error rates, development teams can gain a deeper understanding of how the mock server is being used and identify areas for improvement. These analytics could also assist in optimizing performance and troubleshooting issues

XII. Conclusion

In conclusion, the custom mocking server offers a robust and flexible solution that not only simplifies the process of simulating API endpoints but also empowers development teams to work more efficiently and securely. By seamlessly integrating with the project, the system ensures that sensitive data remains protected, while offering full control over how APIs are defined, managed, and tested. The ability to simulate microservices interactions and interact directly with mock data through Swagger UI adds an extra layer of usability, enhancing the developer experience and reducing dependencies on real services during testing and development.

As the world of API-driven development continues to evolve, the future of this custom solution looks even more promising. With potential improvements like AI-powered mock response generation, support for modern technologies like GraphQL, and sophisticated analytics for deeper insights into API usage, the mock server is poised to become an indispensable tool for teams looking to stay ahead of the curve. These advancements will make it not just a tool for testing, but an intelligent platform that adapts to developers needs, providing dynamic, context-aware simulations that align with real-world scenarios.

By staying adaptable and forward-thinking, this custom mock server can play a critical role in shaping more agile, efficient, and secure development workflows in the years to come.

REFERENCES

[1] C. R. B. de Souza, D. Redmiles, L. T. Cheng, D. Millen, and J. Patterson, "Sometimes you need to see through walls: a field study of application programming interfaces," Proc. 2004 ACM Conf. Computer Supported Cooperative Work, Nov. 2004, pp. 63–71

[2] J. A. Clark, "Application Programming Interface," The SAGE International Encyclopedia of Mass Media and Society, 2020

[3] Isha and A. Sharma, "Automated API Testing," *Proc. 3rd Int. Conf. Inventive Comput. Technol. (ICICT)*, Nov. 2018, doi: 10.1109/ICICT43934.2018.9034254

[4] T. Xin, V. C. Pham, M. Sioutis, Y. Lim, and Y. Tan, "A mock server for ECHONET Lite Web API application development and testing," 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 2022, pp. 786-787, doi: 10.1109/GCCE56475.2022.10014076

[5] A. Soni, V. Ranga, and S. Jadhav, "MockRest—A generic approach for automated mock framework for REST APIs generation," in *Inventive Communication and Computational Technologies: Proceedings of ICICCT 2019*, Springer Singapore, 2020, pp. 237–255, doi: 10.1007/978-981-15-0146-3_24

[6] C. S. Pareek, "From automation to intelligence: Revolutionizing microservices and API testing with AI," Int. J. Res. Appl. Sci. Eng. Technol., vol. 12, no. 11, pp. 716-723, Nov. 2024, doi: 10.22214/ijraset.2024.65198

[7] O. Deineka and L. Bortnik, "Methodology for collecting, processing, storing and classifying data in accordance with SOC2 Type2 requirements," *Computer Systems and Network*, vol. 6, pp. 36-43, Jun. 2024, doi: 10.23939/csn2024.01.036

[8] O. Deineka, O. Harasymchuk, A. Partyka, A. Obshta, and N. Korshun, "Designing data classification and secure store policy according to SOC 2 Type II," *Cybersecurity Providing in Information and Telecommunication Systems 2024*, vol. 3654, pp. 398-409, 2024

[9] R. Chittimalla, "Data Protection in Collaboration: Safeguarding Sensitive Information with Third-Party Stakeholders," Int. J. Sci. Res.(IJSR), vol. 10, no. 2, pp. 1766 1769, Feb. 2021, doi: 10.21275/ES21208181554

[10] J. Seaman, "PCI DSS Applicability," *PCI DSS: An Integrated Data Security Standard Guide*, 1st ed., pp. 195-211, May 2020, doi: 10.1007/978-1-4842-5808-8_7

[11] A. Alpatov, "Simulation server REST API based on the specification for testing web applications," Sovrem. Nauk. Tekhnol. (Modern High Technol.), vol. 2, no. 11, Dec. 2020, doi: 10.17513/snt.38371

[12] C. Hekmatyar, "Utilizing mocking during the development phase," Bachelor's thesis, Information and Communication Technology, Riihimäki, Autumn 2023

[13] [Online]. Available: https://www.postman.com/product/what-is-postman/. [Accessed: Mar. 8, 2025]

[14] [Online]. Available: https://learning.postman.com/docs/design-apis/mock apis/set-up-mock-servers/. [Accessed: Mar. 8, 2025]

[15] [Online]. Available: https://wiremock.org/docs/. [Accessed: Mar. 8, 2025]

[16] [Online]. Available: https://docs.wiremock.io/data-sources/overview. [Accessed: Mar. 8, 2025]

[17] R. Narwal, S. Goyal, and K. Ahuja, "An in-depth study of virtual machines, Docker technology, and performance," *Int. J. Adv. Res. Sci., Commun. Technol.*, vol. 4, no. 6, Apr. 2024, pp. 546, doi: 10.48175/IJARSCT-17670

[18] A. Chandrachood, "API first development: A modern approach to building integrated software systems," Int. J. Sci. Res. (IJSR), vol. 10, no. 10, pp. 1627-1629, Oct. 2021, doi: 10.21275/SR24608141605

[19] Chapke, K. Akant, and P. Chandankhede, "Strategic approaches to modern data management leveraging relational database systems," in Proc. 2024 Int. Conf. Inventive Comput. Technol. (ICICT), Lalitpur, Nepal, 2024, pp. 353-358, doi: 10.1109/ICICT60155.2024.10544431

[20] K. Balasubramanian, "Advanced computing: Best practices for automating image deployment using Docker, cloud, and HPC for AI/ML applications," EPRA Int. J. Res. Dev. (IJRD), Oct. 2024, doi: 10.36713/epra18562

[21] P. S. Emmanni, "Implementing CI / CD Pipelines for Enhanced Efficiency in IT Projects," Int. J. Sci. Res. (IJSR), vol. 9, no. 9, Sep. 2018, doi: 10.21275/SR24402001528

[22] N. Singh, "CI/CD Pipeline for Web Applications," Int. J. Res. Appl. Sci. Eng. Technol., vol. 11, no. 5, pp. 5218-5226, May 2023, doi: 10.22214/ijraset.2023.52867

[23] [Online]. Available: https://docs.scala-lang.org/tour/tour-of-scala.html. [Accessed: Mar. 10, 2025]

[24] [Online]. Available: https://docs.gatling.io/. [Accessed: Mar. 10, 2025]

[25] Baudry, K. Etemadi, S. Fang, Y. Gamage, Y. Liu, Y. Liu, M. Monperrus, J. Ron, A. Silva, and D. Tiwari, "Generative AI to generate test data generators," IEEE Soft., vol. PP, no. 99, pp. 1-9, Nov. 2024, doi: 10.1109/MS.2024.3418570

[26] W. Sampurno, N. M. I. M. Mandennic, and J. R. K. U. B. Jimbaran, "Graph-QL Responsibility Analysis at Integrated Competency Certification Test System Base on Web Service," Lontar Komput. J. Ilm. Teknol. Inf., vol. 11, no. 2, p. 114, Oct. 2020, doi: 10.24843/LKJITI.2020.v11.i02.p05

[27] M. Vesić and N. Kojić, "Comparative analysis of web application performance in case of using REST versus GraphQL," in Proc. Fourth Int. Sci. Conf. ITEMA Recent Adv. Inf. Technol., Tourism, Econ., Manag. and Agric., Jan. 2020, doi: 10.31410/ITEMA.2020.17

[28] A. Van der Hoek, E. Dincel, and N. Medvidovic, "Using service utilization metrics to assess and improve product line architectures," USC-CSE-2001-507, 2001

Graph-Based Approaches to Input-Output Analysis, Trade Flows, and Economic Complexity: A Literature Review

1st Dimitar Peshevski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
dimitar.peshevski@finki.ukim.mk

2nd Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
dimitar.trajanov@finki.ukim.mk

***Abstract*—Economic systems involve interdependent industries, trade networks, and production structures. Traditional input-output (IO) models help analyze these interactions but face limitations in capturing dynamic relationships. Recent advances in network science, economic complexity theory, and knowledge graphs have provided new methods for studying economic structures and trade flows. This review examines the application of network-based approaches in economic analysis, focusing on IO models, trade networks, and causal modeling. It explores how graph-based methods enhance trade and supply chain analysis, improve economic forecasting, and support industrial decision-making. The findings highlight these techniques' potential to address economic modeling challenges, such as data integration and scalability.**

***Index Terms*—economic networks, input-output models, trade networks, economic complexity, network science, knowledge graphs, supply chain analysis, graph-based methods, causal modeling, graph neural networks, industrial decision-making.**

I. Introduction

Economic systems are inherently complex, characterized by intricate relationships between industries, regions, and global trade partners. Understanding these relationships is essential for analyzing production structures, trade flows, and economic resilience. Input-output (IO) models have traditionally been used to study inter-sectoral dependencies, offering valuable insights into economic interactions. However, conventional IO approaches often face limitations in capturing dynamic relationships and network structures. The theory of economic complexity provides an additional perspective by examining how productive capabilities shape economic development. Meanwhile, trade networks highlight the role of diversification and connectivity in fostering growth. The increasing availability of large-scale economic data has created opportunities for leveraging advanced analytical methods, including network science and knowledge graphs, to improve economic modeling and decision-making.

In this paper, we explore recent advancements in network-based approaches to economic analysis, focusing on IO models, trade networks, and economic complexity. We examine the role of graph-based methods in trade and supply chain analysis, the application of knowledge graphs in economic forecasting and manufacturing, and the use of causal modeling in economic systems. By synthesizing these developments, the review aims to highlight key contributions, identify research gaps, and outline future directions for integrating network science and AI-driven methodologies into economic studies.

II. Literature Review

In this section, we provide a comprehensive overview of the methodologies used to analyze economic networks, trade, and supply chains, with a particular emphasis on graph-based approaches. Table I summarizes key methodologies from the existing literature, categorizing them based on analytical techniques and application areas. The papers are grouped according to their primary methodological focus and application domain. Additionally, Figure 1 presents a more detailed breakdown, illustrating the relationships between methodologies and application areas using a Sankey plot, illustrating the distribution of papers employing each approach.

The review is organized into six subsections, each dedicated to a specific methodological framework. We begin by discussing input-output models and their role in understanding economic networks. Next, we examine economic complexity measures and their applications in analyzing trade networks. The third subsection focuses on graph-based methods specifically used for studying trade and supply chain structures. We then explore network analysis techniques applied to international trade data. Following this, we review the use of knowledge graphs in trade and economic studies. Finally, we discuss causal modeling approaches and their relevance in examining trade and economic systems. Throughout this review, we emphasize the importance of graph-based methodologies in capturing the structure and dynamics of economic interactions.

A. Input-Output Models and Economic Networks

Input-output (IO) models are widely used to analyze economic interactions between sectors and regions. Traditional IO models, however, often face challenges related to data granularity and integration. To address these limitations, researchers

TABLE I
COMPARISON OF THE PRIMARY METHODOLOGIES AND APPLICATION AREAS USED IN THE REVIEWED LITERATURE

Papers	Methodology	Application Area
[1], [4], [5], [10], [26], [37], [41], [43], [44], [49]	Network Analysis – Graph Theory, Centrality Measures, Graphlet Analysis, Algebraic Transformations	Input-Output Models & Economic Structure – Economic Networks, Trade Networks, Production Systems
[14], [15], [17]–[19], [42]	Economic Complexity Methods – Product Space, Economic Fitness, Diversification Metrics	Economic Complexity & Trade Patterns – Industrial Diversification, Economic Growth, International Trade
[29], [33], [39]	Probabilistic & Stochastic Models – Random Walks, Markov Models, Exponential Random Graph Models (ERGM)	Trade Network Evolution & Global Trade Patterns – Trade Flow, Market Dynamics, Bilateral Trade Relations
[7], [21], [31], [32], [35], [36], [38], [51]	Knowledge Graphs & Embeddings – Knowledge Representation, Graph Embeddings, Knowledge Graph Reasoning	Supply Chain & Industrial Optimization – Supply Chain Risk Management, Logistics, Industrial Resilience
[23]–[25], [27], [28], [45], [46], [52]	Graph Neural Networks (GNNs) – Graph Convolutional Networks, Graph Auto-Encoders, Graph Attention Networks	Supply Chain Risk Management – Disruption Prediction, Resilience Strategies, Uncertainty Modeling
[11], [16], [22], [40]	Econometric Models – Panel Data Analysis, Gravity Models, Regression Analysis	Trade Prediction & Export Diversification – Trade Forecasting, Market Entry Strategies, Competitiveness
[6], [8], [9], [12], [20], [30], [48]	Causal Inference – Granger Causality, Structural Causal Models, Bayesian Networks	Causal Relationships in Trade & Economic Growth – Policy Evaluation, Circular Economy, Market Dependencies
[2], [3], [13], [34], [47], [50]	Semantic Web & Ontology-Based Modeling – Knowledge Graphs, Ontologies, Knowledge Representation	Manufacturing & Industrial Knowledge Representation – Industry 4.0, Process Optimization, Smart Manufacturing

have explored network-based approaches that represent IO systems as interconnected structures. This transformation facilitates the incorporation of empirical flow data without requiring additional estimations [37]. Another development in IO analysis is the use of Physical Input-Output Tables (PIOTs), which track material flows within economies. Modular bottom-up methodologies have been proposed to enhance the accuracy and automation of PIOT construction [43]. Furthermore, large-scale databases such as the World Input-Output Database (WIOD) provide standardized IO data, enabling cross-country economic comparisons and policy analysis [5], [41].

Network analysis has contributed to a deeper understanding of economic structures, production linkages, and aggregate fluctuations. Studies on production networks have examined how local shocks propagate through economies, influencing overall stability and performance [4]. Graph-based methods have been applied to sectoral interactions, identifying structural patterns that are consistent across different economies [26]. The integration of network analysis tools with IO models allows for the measurement of sectoral interdependencies using various network metrics [49]. Additionally, graphlet-based approaches have been utilized for production optimization in industrial settings, demonstrating their potential for improving efficiency and decision-making [44]. These advancements highlight the role of network-based methodologies in modern IO analysis and their relevance for economic planning and resource allocation.

B. Economic Complexity and Trade Networks

The theory of economic complexity provides a framework for understanding how productive knowledge is embedded in economies. Economic growth is linked to a country's ability to diversify and produce complex goods, which depends on the underlying capabilities required for production [15], [17], [42]. The concept of the "product space" describes how economies develop by shifting toward more sophisticated products. Countries located in highly connected regions of the product space are more likely to achieve sustained growth [19]. Research indicates that economies with greater productive knowledge tend to experience higher income levels and long-term economic expansion [18].

Trade networks play a significant role in economic complexity by influencing diversification strategies. The relationship between export diversification and economic performance suggests that countries exporting a broad range of products, particularly manufactured goods, tend to experience stronger economic growth [11], [14]. Empirical studies support the view that diversification contributes to economic development by facilitating knowledge spillovers and enhancing competitiveness in global markets [16]. Additionally, advanced graph-learning techniques applied to the product space have improved methods for recommending industrial diversification strategies. These approaches demonstrate the value of graph-learning techniques in informing economic policy decisions [7], [32], [40].

C. Graph-Based Methods for Trade and Supply Chain Analysis

Graph-based methods are increasingly used to analyze trade and supply chain networks, providing insights into structural dependencies, risk propagation, and predictive modeling. One prominent approach is the application of Markov chain models to world input-output networks, which has been used to assess economic globalization, systemic risks, and monetary flow dynamics [29], [33]. These methods help identify key economic hubs and pathways through which economic disturbances

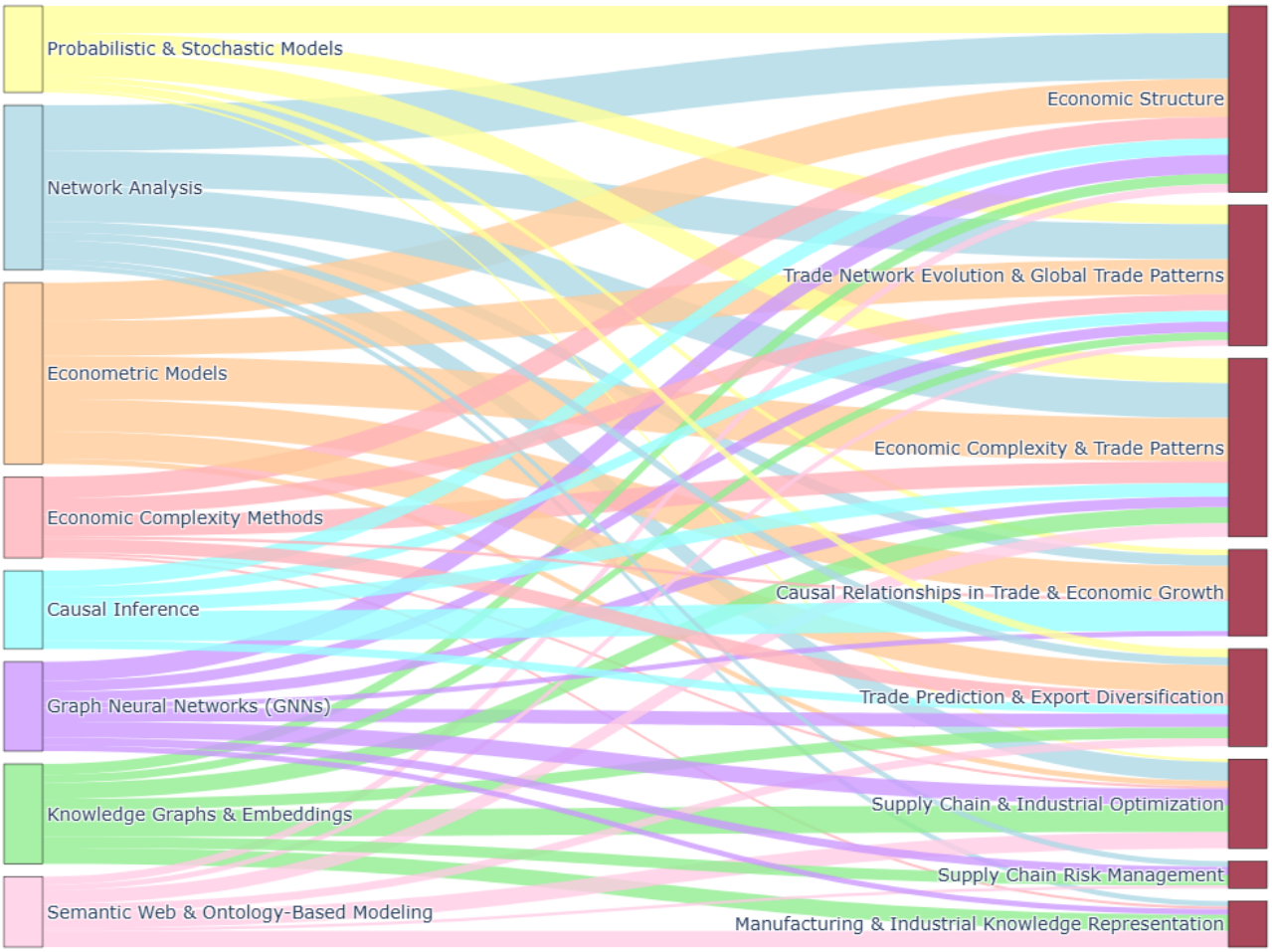


Fig. 1. A Sankey plot illustrating the distribution and relationships between the methodologies and their corresponding application areas, as identified in the reviewed literature. The plot visually represents the distribution of papers employing each methodology across various economic domains.

can spread. Additionally, machine learning models, including Graph Neural Networks (GNNs) [25], have been applied to trade networks to address causal inference challenges and analyze disruptions due to external shocks. GNNs have been used to adjust for high-dimensional network confounding, making them effective for studying economic interdependencies [28]. Furthermore, specialized AI-based methods, such as the Artificial Intelligence Network Explanation of Trade (AINET), have been developed to model trade networks and predict shifts in trade flows due to outlier events like tariffs, natural disasters, and trade wars [24]. Also, a recent development using GNNs is the use of Gravity-informed Graph Autoencoders (GGAE) for trade volume prediction, as proposed in [27]. These approaches enhance the ability to understand global trade dynamics and inform policy decisions.

In supply chain analysis, graph-based methodologies have been instrumental in improving risk management and operational efficiency. GNNs have been applied to supply chain networks to enhance industry classification, link prediction, and risk identification [2], [45], [46]. These models utilize the structural properties of supply chains to detect vulnerabilities and assess the resilience of production systems. Additionally, domain knowledge graphs have been used to integrate multi-source data for enhanced risk assessment and decision-making [23], [52]. By combining graph-based techniques with supply chain analytics, researchers have developed more accurate forecasting models, proactive risk mitigation strategies, and improved classification of economic entities. These advancements highlight the growing importance of network science in trade and supply chain studies, demonstrating its potential for enhancing economic resilience and policy formulation.

D. Network Analysis of International Trade

Network analysis provides a structured approach to examining international trade relationships, capturing the structural, economic, and geopolitical factors that shape trade flows. Studies utilizing exponential random graph models have highlighted the influence of geography, political ties, and

economic strength in forming trade connections, revealing the underlying patterns that drive global trade interactions [39]. Similarly, social network analysis techniques applied to export networks have demonstrated the role of geographic proximity and centrality in shaping international trade, particularly in response to external shocks like the COVID-19 pandemic [10]. The long-term structural properties of the International Trade Network (ITN) indicate a scale-invariant nature, emphasizing persistent global trade patterns over decades [1].

Beyond structural analysis, network-based approaches have been used to investigate causal relationships between trade and other economic activities. Studies integrating trade networks with tourism data suggest that increased international human interactions can stimulate trade flows, highlighting the interplay between mobility and commerce [22]. Additionally, the application of Granger causality in trade and patenting networks indicates a feedback loop between trade participation and international innovation diffusion, where patenting activity influences export behavior [30]. These findings underscore the interconnectedness of global trade and other economic networks, demonstrating the utility of network analysis in understanding trade dynamics.

E. Knowledge Graphs in Trade, Economy, and Manufacturing

Knowledge graphs (KGs) have been increasingly applied to trade and economic analysis to enhance predictive accuracy and capture complex relationships. In international trade, KG-based models have been used to improve the prediction of trade flows by leveraging graph neural networks and embedding techniques, outperforming traditional statistical methods [36], [38]. KGs have also been integrated with macroeconomic data, incorporating alternative big data sources to refine economic forecasting models [51]. Furthermore, knowledge graph embeddings have been shown to enhance the accuracy of trade flow predictions by identifying missing relationships within trade networks [31], [35]. These applications provide policymakers and businesses with a structured approach to understanding international trade dynamics and anticipating economic changes.

In manufacturing and production, KGs play a role in advancing Industry 4.0 by optimizing process planning and resource management. Studies have explored KG applications in intelligent manufacturing, including process knowledge representation, automated knowledge extraction, and production planning [3], [21], [34], [47]. The integration of KGs with semantic web technologies has enabled real-time monitoring and optimization of industrial processes [50]. Additionally, KGs have facilitated the construction of machining process knowledge bases, improving efficiency in intelligent process design [13]. These advancements demonstrate the potential of knowledge graphs in enhancing decision-making and automation across trade, economic forecasting, and manufacturing sectors.

F. Causal Modeling in Trade and Economic Systems

Causal modeling plays a crucial role in understanding the complex interdependencies within trade and economic systems. The application of causal inference methods in economic time series models has been widely explored, particularly in relation to predictability, exogeneity, and model construction [9]. These approaches align with econometric techniques such as Wiener–Granger causality and structural equation modeling, enabling researchers to identify directional relationships between economic variables. Similarly, in the context of knowledge-based economies, causal modeling has been used to analyze the impact of innovation systems, human resources, and information technology on economic competitiveness, employing structural modeling techniques to assess causality between key economic factors [6].

Recent advancements in causal modeling extend to areas such as circular economies and financial markets. Circular and cumulative causation models illustrate feedback loops in economic growth, demonstrating how trade can contribute to circular economic development by influencing resource allocation and environmental sustainability [8]. Machine learning techniques have also been integrated into causal analysis, with studies leveraging advanced statistical models to improve predictions in financial and economic data [20], [48]. Furthermore, causal network maps have been employed to analyze urban circular economies, providing a structured visualization of economic drivers and their interrelationships [12]. These diverse applications underscore the significance of causal modeling in shaping policy decisions and optimizing economic outcomes.

III. Discussion and Future Directions

The reviewed literature highlights the growing significance of network-based methodologies, knowledge graphs, and AI-driven techniques in economic analysis. Input-output models have evolved through the integration of network science, enhancing the understanding of economic interdependencies and shock propagation. Similarly, economic complexity theory and trade networks emphasize the role of diversification and connectivity in driving growth. Graph-based methods and knowledge graphs have improved trade and supply chain analysis by providing more accurate predictions and risk assessments. Moreover, causal modeling has facilitated deeper insights into economic relationships, extending beyond traditional econometric techniques. Despite these advancements, several research gaps remain, particularly in the scalability of network models, integration of multi-source data, and the interpretability of AI-driven economic insights.

Future research should focus on refining network-based approaches by improving data availability and standardization across economic datasets. The application of knowledge graphs in trade and manufacturing requires further development to enhance automated reasoning and decision-making. AI techniques, including Graph Neural Networks and causal inference methods, should be expanded to capture non-linear

and dynamic economic relationships more effectively. Additionally, interdisciplinary studies combining network science, AI, and policy analysis could provide a more comprehensive framework for economic forecasting and risk management. Addressing these gaps will contribute to more robust economic models, facilitating better policy decisions and industrial strategies in an increasingly interconnected global economy.

IV. Conclusion

This review highlights the increasing role of network-based methodologies in economic analysis, particularly in input-output modeling, trade networks, and economic complexity. The integration of network science and graph-based techniques has provided deeper insights into sectoral interdependencies, trade diversification, and supply chain dynamics. Knowledge graphs have enhanced economic forecasting and industrial planning, while causal modeling has improved the understanding of economic relationships. These advancements address limitations in traditional economic models by leveraging large-scale data and computational techniques to capture complex interactions within economic systems.

For policymakers, economists, and industry practitioners, these findings underscore the importance of adopting network-based approaches for more informed decision-making. The ability to analyze economic resilience, predict trade disruptions, and optimize industrial processes can enhance policy effectiveness and business strategies. However, challenges remain in terms of data integration, scalability, and model interpretability. Future research should focus on refining these methodologies to improve their applicability and reliability. As economic systems become increasingly interconnected, network science will continue to play a critical role in advancing economic analysis and decision-making.

References

[1] K Bhattacharya, G Mukherjee, and SS Manna. The international trade network. In *Econophysics of Markets and Business Networks: Proceedings of the Econophys-Kolkata III*, pages 139–147. Springer, 2007.

[2] Nils Brockmann, Edward Elson Kosasih, and Alexandra Brintrup. Supply chain link prediction on uncertain knowledge graph. *ACM SIGKDD Explorations Newsletter*, 24(2):124–130, 2022.

[3] Georg Buchgeher, David Gabauer, Jorge Martinez-Gil, and Lisa Ehrlinger. Knowledge graphs in manufacturing and production: a systematic literature review. *IEEE Access*, 9:55537–55554, 2021.

[4] Vasco M Carvalho. From micro to macro via production networks. *Journal of Economic Perspectives*, 28(4):23–48, 2014.

[5] Federica Cerina, Zhen Zhu, Alessandro Chessa, and Massimo Riccaboni. World input-output network. *PloS one*, 10(7):e0134025, 2015.

[6] Chih-Kai Chen. Causal modeling of knowledge-based economy. *Management Decision*, 46(3):501–514, 2008.

[7] Kévin Cortial, Adélaïde Albouy-Kissi, and Frédéric Chausse. Product space clustering with graph learning for diversifying industrial production. *Applied Sciences*, 14(7):2833, 2024.

[8] Kieran P Donaghy. A circular economy model of economic growth with circular and cumulative causation and trade. *Networks and Spatial Economics*, 22(3):461–488, 2022.

[9] John Geweke. Inference and causality in economic time series models. *Handbook of econometrics*, 2:1101–1144, 1984.

[10] Dilek Gönçer-Demiral and Meltem İnce-Yenilmez. Network analysis of international export pattern. *Social Network Analysis and Mining*, 12(1):156, 2022.

[11] Raul Gouvea, Gautam Vora, et al. Reassessing export diversification strategies: a cross-country comparison. *Modern Economy*, 6(01):96, 2015.

[12] Ivan Henderson V Gue, Raymond R Tan, and Aristotle T Ubando. Causal network maps of urban circular economies. *Clean Technologies and Environmental Policy*, pages 1–12, 2022.

[13] Liang Guo, Fu Yan, Tian Li, Tao Yang, and Yuqian Lu. An automatic method for constructing machining process knowledge base from knowledge graph. *Robotics and Computer-Integrated Manufacturing*, 73:102222, 2022.

[14] Ricardo Hausmann and César Hidalgo. Country diversification, product ubiquity, and economic divergence. 2010.

[15] Ricardo Hausmann, César A Hidalgo, Sebastián Bustos, Michele Coscia, and Alexander Simoes. *The atlas of economic complexity: Mapping paths to prosperity*. Mit Press, 2014.

[16] Dierk Herzer and Felicitas Nowak-Lehnmann D. What does export diversification do for growth? an econometric analysis. *Applied economics*, 38(15):1825–1838, 2006.

[17] Josiah Hickson. The atlas of economic complexity: A review. *Newcastle Business School Student Journal*, 1(1):27–33, 2017.

[18] César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26):10570–10575, 2009.

[19] César A Hidalgo, Bailey Klinger, A-L Barabási, and Ricardo Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

[20] Andrew Horton. Causal economic machine learning (ceml):" human ai". *AI*, 5(4), 2024.

[21] Jiaqing Huang, Xiaojun Zhang, Xinyu Hu, Hongyan Wu, and Wen Zhang. Construction and teaching application of knowledge graph in the field of international trade. In *Proceedings of the International Conference on Decision Science & Management*, pages 270–275, 2024.

[22] Kiyong Keum. International tourism and trade flows: A causality analysis using panel data. *Tourism Economics*, 17(5):949–962, 2011.

[23] Edward Elson Kosasih, Fabrizio Margaroli, Simone Gelli, Ajmal Aziz, Nick Wildgoose, and Alexandra Brintrup. Towards knowledge graph reasoning for supply chain risk management using graph neural networks. *International Journal of Production Research*, 62(15):5596–5612, 2024.

[24] Michael P Leung and Pantelis Loupos. Graph neural networks for causal inference under network confounding. *arXiv preprint arXiv:2211.07823*, 2022.

[25] Jovan Manchev, Mirsolav Mirchev, and Igor Mishkovski. Classification of companies using graph neural networks. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 1087–1092. IEEE, 2024.

[26] James McNerney. Network properties of economic input-output networks. 2009.

[27] Naoto Minakawa, Kiyoshi Izumi, and Hiroki Sakaji. Bilateral trade flow prediction by gravity-informed graph auto-encoder. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2327–2332. IEEE, 2022.

[28] Anderson Monken, Flora Haberkorn, Munisamy Gopinath, Laura Freeman, and Feras A Batarseh. Graph neural networks for modeling causality in international trade. In *The international FLAIRS conference proceedings*, volume 34, 2021.

[29] Vahid Moosavi and Giulio Isacchini. A markovian model of evolving world input-output network. *PloS one*, 12(10):e0186746, 2017.

[30] Bachar Moussa and Nikos C Varsakelis. Causality between international trade and international patenting: A combination of network analysis and granger causality. *Atlantic Economic Journal*, 50(1):9–26, 2022.

[31] Durgesh Nandini, Simon Blöthner, Mirco Schoenfeld, and Mario Larch. Multidimensional knowledge graph embeddings for international trade flow analysis. *arXiv preprint arXiv:2410.19835*, 2024.

[32] Arnault Pachot, Adélaïde Albouy-Kissi, Benjamin Albouy-Kissi, and Frédéric Chausse. Production2vec: a hybrid recommender system combining semantic and product complexity approach to improve industrial resiliency. In *2021 2nd international conference on artificial intelligence and information systems*, pages 1–6, 2021.

[33] Carlo Piccardi, Massimo Riccaboni, Lucia Tajoli, and Zhen Zhu. Random walks on the world input–output network. *Journal of Complex Networks*, 6(2):187–205, 2018.

[34] Jing Quan. Visualization and analysis model of industrial economy status and development based on knowledge graph and deep neural network. *Computational Intelligence and Neuroscience*, 2022(1):7008093, 2022.

[35] Diego Rincon-Yanez, Amira Mouakher, and Sabrina Senatore. Enhancing downstream tasks in knowledge graphs embeddings: a complement graph-based approach applied to bilateral trade. *Procedia Computer Science*, 225:3692–3700, 2023.

[36] Diego Rincon-Yanez, Chahinez Ounoughi, Bassem Sellami, Tarmo Kalvet, Marek Tiits, Sabrina Senatore, and Sadok Ben Yahia. Accurate prediction of international trade flows: Leveraging knowledge graphs and their embeddings. *Journal of King Saud University-Computer and Information Sciences*, 35(10):101789, 2023.

[37] João Rodrigues, Alexandra Marques, Richard Wood, and Arnold Tukker. A network approach for assembling and linking input–output models. *Economic Systems Research*, 28(4):518–538, 2016.

[38] Bassem Sellami, Chahinez Ounoughi, Tarmo Kalvet, Marek Tiits, and Diego Rincon-Yanez. Harnessing graph neural networks to predict international trade flows. *Big Data and Cognitive Computing*, 8(6):65, 2024.

[39] Amin Setayesh, Zhivar Sourati Hassan Zadeh, and Behnam Bahrak. Analysis of the global trade network using exponential random graph models. *Applied Network Science*, 7(1):38, 2022.

[40] Viktor Stojkoski, Philipp Koch, Eva Coll, and César A Hidalgo. Estimating digital product trade through corporate revenue data. *Nature Communications*, 15(1):5262, 2024.

[41] Marcel P Timmer, Erik Dietzenbacher, Bart Los, Robert Stehrer, and Gaaitzen J De Vries. An illustrated user guide to the world input–output database: the case of global automotive production. *Review of International Economics*, 23(3):575–605, 2015.

[42] Zoran Utkovski, Melanie F Pradier, Viktor Stojkoski, Fernando Perez-Cruz, and Ljupco Kocarev. Economic complexity unfolded: Interpretable model for the productive structure of economies. *PloS one*, 13(8):e0200822, 2018.

[43] Liz Wachs and Shweta Singh. A modular bottom-up approach for constructing physical input–output tables (piots) based on process engineering models. *Journal of Economic Structures*, 7:1–24, 2018.

[44] Zun Wang, Yongming Han, Chengfei Li, Zhiqiang Geng, and Jinzhen Fan. Input-output networks considering graphlet-based analysis for production optimization: Application in ethylene plants. *Journal of Cleaner Production*, 278:123955, 2021.

[45] Azmine Toushik Wasi, MD Islam, Adipto Raihan Akib, and Mahathir Mohammad Bappy. Graph neural networks in supply chain analytics and optimization: Concepts, perspectives, dataset and benchmarks. *arXiv preprint arXiv:2411.08550*, 2024.

[46] Desheng Wu, Quanbin Wang, and David L Olson. Industry classification based on supply chain network information using graph neural networks. *Applied Soft Computing*, 132:109849, 2023.

[47] Youzi Xiao, Shuai Zheng, Jiancheng Shi, Xiaodong Du, and Jun Hong. Knowledge graph-based manufacturing process planning: A state-of-the-art review. *Journal of Manufacturing Systems*, 70:417–435, 2023.

[48] Lei Xu. Machine learning and causal analyses for modeling financial and economic data. In *Applied Informatics*, volume 5, page 11. Springer, 2018.

[49] Ming Xu and Sai Liang. Input–output networks offer new insights of economic structure. *Physica A: Statistical Mechanics and its Applications*, 527:121178, 2019.

[50] Muhammad Yahya, John G Breslin, and Muhammad Intizar Ali. Semantic web and knowledge graphs for industry 4.0. *Applied Sciences*, 11(11):5110, 2021.

[51] Yucheng Yang, Yue Pang, Guanhua Huang, et al. The knowledge graph for macroeconomic analysis with alternative big data. *arXiv preprint arXiv:2010.05172*, 2020.

[52] Wanyue Zhang, Yonggang Liu, Lihong Jiang, Nazaraf Shah, Xiang Fei, and Hongming Cai. The construction of a domain knowledge graph and its application in supply chain risk analysis. In *Advances in E-Business Engineering for Ubiquitous Computing: Proceedings of the 16th International Conference on e-Business Engineering (ICEBE 2019)*, pages 464–478. Springer, 2020.

VulnerSec: A Flexible, Automated and Open-Source Cybersecurity Framework

Evgenija Krajchevska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
evgenija.krajchevska@finki.ukim.mk

Vojdan Kjorveziroski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
vojdan.kjorveziroski@finki.ukim.mk

***Abstract*—Hands-on cybersecurity training is crucial for developing practical skills, yet existing platforms often present limitations such as being closed-source, requiring complex setup or relying on outdated vulnerabilities that become unusable over time. To tackle these issues, we introduce an open-source framework designed to create flexible, customizable and reproducible cybersecurity scenarios. It provides a modular approach, allowing deployment of vulnerable environments with minimal setup overhead, and addresses legacy software vulnerabilities making them accessible even on modern systems. The decoupling of the provisioning and deployment processes ensures greater adaptability, allowing the same configurations to be applied across diverse infrastructures with minimal adjustments while widely adopted tools are used to simplify these processes. The framework was validated in a university cybersecurity course where students participated in vulnerability discovery and exploitation exercises. Future work will primarily focus on expanding the vulnerability set, implementing profile-based configurations that provide customizable scenarios based on complexity levels as well as implementation of network monitoring and generation of datasets for the improvement of intrusion detection and threat mitigation processes.**

***Index Terms*—hands-on cybersecurity, software vulnerabilities, framework, automation, virtualization**

I. Introduction

With the rising demand for IT professionals proficient in ever-changing technologies and tools, it is evident that practical work in computer science has become essential, even in introductory courses accommodating learners of diverse backgrounds and varying skill levels [1]. Hands-on exercises and project-based learning constitute an important aspect of both formal and informal education, as they immerse students in real-world tasks and foster active engagement with course material. Extensive studies show that interactive, hands-on activities boost engagement and performance in science and cybersecurity courses alike [2] [3].

When it comes to cybersecurity, research highlights that structured cybersecurity exercises improve both technical proficiency and decision making skills by immersing learners in realistic threat scenarios [4]. Challenge-based learning, especially in the form of Capture The Flag (CTF) exercises, can also be seen as an effective methodology in cybersecurity education, demonstrating improvements in practical skills, motivation, and participation of students when faced with real-world scenarios [5].

The main obstacle when organizing hands-on exercises is that their planning and infrastructure setup takes too much time from the educators. This is especially the case with cybersecurity since the integrity of the entire system must be considered. Unlike general computing tasks, these exercises require secure isolated environments to avoid compromising institutional networks and demand constant maintenance and updates to keep pace with software releases. Additionally, each student requires their own personalized environment, further complicating things. Providing every learner with an isolated virtual machine instantiates what Cantelli-Forti and Colajanni describe as the stateful paradigm, where the environment preserves interaction state across sessions, thereby intensifying both setup and maintenance overhead [6]. Some of these challenges can be alleviated by using proprietary services and platforms, albeit they introduce their own drawbacks.

One common limitation of commercial solutions is that subscription fees can be prohibitively expensive for institutions with limited budgets, particularly in developing countries, while the lack of source code access makes it difficult for instructors or researchers to fully understand or adapt the underlying security controls and architectures. Furthermore, most proprietary offerings provide only a fixed set of scenarios or challenges, preventing educators from personalizing content to specific learning objectives or incorporating novel vulnerabilities. By contrast, open-source solutions are freely available to both instructors and researchers, stimulating swift iteration and collaborative development through community-driven repositories. Such solutions often embrace modular design and integrate seamlessly with a variety of hypervisors, operating systems, or cloud platforms, making it easier to keep pace with evolving requirements.

Aside from being a valuable asset in educational practices, these frameworks play a critical role in cybersecurity research, specifically in studying real-world attack behaviors and defense strategies. Due to their isolated nature, vulnerable environments are often used to observe attack vectors, analyze vulnerabilities, and simulate realistic attack-defense scenarios. Multiple recent efforts highlight the importance of novel research in this domain. Simulation frameworks remain an

effective way to model multi-stage attacks and observe the interplay between complex network vulnerabilities, security devices, and traffic patterns [7]. Modeling frameworks take into account the whole lifecycle of a cyber defense, starting from proactive prevention and detection all the way to thorough incident analysis and recovery which allows researchers to evaluate how various defensive measures perform under realistic threat conditions [8]. By simulating various threat vectors in controlled yet scalable testbeds, cybersecurity frameworks enable researchers to evaluate different defense mechanisms without putting production environments at risk.

The goal of this paper is to present an open-source framework for generating and customizing virtual environments to support hands-on cybersecurity exercises. By simplifying infrastructure and tooling, instructors can deploy vulnerable software with greater flexibility, relying on standardized tools with few prerequisites. Additionally, a container-based approach is adopted to overcome scenario rot and ensure that legacy vulnerabilities remain usable for teaching and research, regardless of host system updates. By offering this solution under a permissive open-source license, we aim to encourage collaboration, adaptability and sustainable use for educators and researchers alike.

The remainder of the paper is structured as follows. Section II discusses existing platforms and scenarios for hands-on exercises, highlighting related work in the field. Section III presents our proposed requirements for a flexible framework, focusing on core aspects such as: customizability, openness, sustainability and ease of use. In Section IV, we introduce the technical architecture and design choices of the presented solution, illustrating how it meets the requirements defined earlier. Section V describes the practical validation of our approach in an introductory cybersecurity course, outlining the deployment process and initial results. Finally, Section VI concludes the paper by summarizing key findings and suggesting directions for future work.

II. Related Work

Various platforms have emerged for integrating vulnerable virtual machines into cybersecurity education, each providing an isolated environment in which students can hone their practical skills. These solutions, frequently adopted in university courses or Capture The Flag (CTF) competitions, enable learners to engage directly with security threats in a controlled setting.

Recent research has explored diverse strategies for the deployment of cybersecurity labs. Russo et al. introduce a gamified cybersecurity education program structured around progressive engagement levels [9]. The entire infrastructure is self-hosted within the faculty's network, ensuring full control over the environment and security configuration. It simulates enterprise-like networks, including firewalls, DNS servers, and client systems, where students individually defend against scripted attacks. While this approach provides high realism and full control over the environment, it introduces remarkable challenges related to manual setup, scalability limitations and resource-intensive maintenance.

In contrast, the CyberArena framework focuses on a scalable, cost-effective architecture utilizing the cloud [10]. Students remotely access isolated lab environments through a browser-based interface while their assessments are carried out through automated scripts. Instructors can add new challenges with the required software, data, user configurations, and automated scripts needed for lab interaction. Proper setup including all necessary dependencies and security settings can be time-intensive. More recently, Nelson et al. introduced DOJO [11], an open-source platform for applied cybersecurity education which incorporates both self-hosted and centrally-hosted deployment models, easing the infrastructure burden for educators. It addresses numerous aspects of lab creation, such as improved accessibility by allowing students to interact with challenges directly through browser-integrated tools as well as providing scalable and reproducible challenge environments by leveraging Docker containers while isolating students' activities. However, it faces constraints in customizing more complex scenarios, particularly those involving legacy vulnerabilities or kernel-level configurations, which often require additional manual setup.

We next examine a selection of well-known platforms that illustrate current practices in the area. One such platform is SecGen [12], an open-source software to automatically provision virtual machines with custom vulnerabilities. Its key-strength lies in the ability to define reproducible scenario-based challenges using a declarative approach. Another open-source project is Vulhub [13], which leverages a container-based approach in the deployment of vulnerabilities, allowing minimal overhead and rapid deployment. It is commonly included as part of cyber ranges developed within this area [14]. Vulnhub [15], by contrast, is not a framework but a community-driven repository hosting a variety of pre-built vulnerable virtual machines that mimic real world security flaws. These machines are frequently used in educational settings to develop penetration testing skills through simulated scenarios. Finally, some platforms deliver "Cybersecurity Labs as a Service", such as CLaaS [16], which provides remote access to pre-configured cybersecurity labs through a web interface, eliminating the burden of local infrastructure management.

Despite the utility of these platforms, each exhibits certain drawbacks that can limit their efficiency in academic contexts. Some rely on older frameworks and require complex scripting, others offer only fixed scenarios without providing customizable environments and a few like CyberArena or CLaaS demand cloud-based infrastructures, potentially limiting accessibility. Moreover, very few solutions address both modern containerized vulnerabilities and legacy exploits that may fail on recent operating systems. Table I outlines the main features and limitations of these solutions, including their deployment models, open-source accessibility, level of customization and hosting requirements.

TABLE I
FEATURE COMPARISON OF CYBERSECURITY TRAINING PLATFORMS

Feature	SecGen	Vulhub	Vulnhub	CLaaS
Type	VM-based scenarios	Container-based scenarios	VM Repository	Cloud-based VM scenarios
Scenario Setup	Puppet	Docker-compose	None (Prebuilt)	VMware/ESXi
Runtime Environment	Virtual machines	Containers	Virtual machines	Virtual machines
Customizable Scenarios	Yes	Yes	No	No
Open-source	Yes	Yes	No	No
Self-hosted	Yes	Yes	Yes	No

III. REQUIREMENTS FOR A FLEXIBLE FRAMEWORK FOR HANDS-ON SCENARIOS

To design an effective framework for hands-on cybersecurity, it is crucial to establish a set of core principles that guarantee adaptability and ease of use. Many existing solutions, as discussed previously, address certain aspects of cybersecurity but often lack the flexibility needed to become truly sustainable.

A. Functional Requirement 1: Scenario Customization

The most significant functional requirement is scenario customization. A well-structured cybersecurity framework must allow users to define configurations dynamically. Generating the challenges using such an approach assures that exercises remain engaging and educationally valuable. Additionally, a cybersecurity framework should be scalable and adaptable, allowing instructors and researchers alike to modify and extend scenarios to accommodate evolving objectives. Hands-on environments must also be resilient to obsolescence. Without proper design considerations, lab scenarios risk becoming outdated, limiting their educational effectiveness. Properly configured environments should not take too much time off of users' hands, as they should minimize complexity and overhead in a way that supports managing and deploying scenarios with ease.

B. Functional Requirement 2: Open-Source Access

An equally important functional requirement is for the platform's source code to be publicly available under a permissive open-source license. A wide range of cybersecurity training solutions have already been developed, yet a common limitation among them is their closed-source nature, restricting the ability to modify or extend their features beyond their original design. This is particularly evident in frameworks developed exclusively for specific institutions, adapted to meet the internal requirements of a single faculty or training program. In addition, an increasing number of commercial cyber ranges operate on a proprietary, subscription-based model, offering training as an online service to individuals seeking to strengthen their skills [17]. By comparison, a solid cybersecurity lab framework should be open-source, inspiring educators and researchers to adjust scenarios, collaborate across institutions and shape scenarios to meet their requirements. Open-source solutions promote long-term sustainability, as they enable a shared development model where improvements and new challenges can be contributed by a broader academic and professional community. By encouraging collaboration among instructors, such frameworks not only enhance the quality and diversity of cybersecurity scenarios but also ensure that training environments remain up to date and widely accessible over time.

C. Functional Requirement 3: Long-Term Sustainability

Cyber-security labs often suffer scenario rot, where exploits break because of software updates, missing dependencies, or vanished resources. We define the third functional requirement as the ability to handle different vulnerability types, including legacy ones. For a framework to be long-term sustainable, it must guarantee that both the scenario and its execution environment remain reproducible over time. Support for multiple diverse runtime environments, each providing isolation features, can prevent compatibility issues and maintain accessibility to older vulnerabilities. A one-size-fits-all approach is insufficient when dealing with a broad range of vulnerabilities, often requiring a flexible deployment strategy, allowing even complicated vulnerability scenarios involving multiple components to be deployed. This approach mitigates dependency issues and preserves older vulnerabilities for continued use. By incorporating support for multiple runtime environments, such a framework offers educators and researchers the flexibility to customize scenarios to suit a variety of necessities.

D. Functional Requirement 4: Ease of Use

Lastly, a functional requirement that should not be overlooked is ease of use. Ideally, frameworks should be straightforward to configure without delving too much into technical details, enabling even educators or researchers who may lack technical skills to navigate the platform with ease. At the same time, the framework must guarantee strict inter-user isolation and protect the host infrastructure by providing practical security safeguards, so students cannot break out of their assigned virtual machines or containers while instructors retain a seamless workflow. The setup process should be intuitive and require minimal manual intervention, relying on automated workflows and widely adopted tools. By simplifying the deployment process the framework remains accessible across diverse educational environments, reducing the technical burden and permitting instructors to focus on various aspects of cybersecurity training and research.

The following section presents a framework that embodies these principles, delivering a scalable and adaptable solution for hands-on cyber security training.

TABLE II
OVERVIEW OF EXAMPLE VULNERABILITIES IN VULNERSEC

Vulnerability Name	Decription	Vulnerability Type	Runtime Environment
Crackable User Account	Weak user credentials	Weak password	Virtual machine
SSH Root Login	SSH root access enabled	Misconfiguration	Virtual machine
NFS Root Share	NFS root export	Misconfiguration	Virtual machine
Netcat Backdoor Chroot Escape	Netcat in chroot environment	Misconfiguration, Backdoor	Virtual machine
UnrealIRC 3.2	Vulnerable IRC Server	Backdoor	Container
ProFTP 1.3.3c	Vulnerable FTP Server	Backdoor	Virtual machine
VsFTP 2.3.4	Vulnerable FTP Server	Backdoor	Virtual machine
Wordpress 3.x	Outdated, vulnerable CMS	Web Application Exploit	Virtual machine
Grafana 8.0.0	Vulnerable Grafana Version	Path Traversal	Container

IV. THE VULNERSEC PLATFORM FOR HANDS-ON SCENARIOS

For a platform to be flexible, it must be capable of adapting easily to various teaching requirements and technical constraints. In alignment with the importance of scenario customization, our proposed framework, VulnerSec, addresses the need for enhanced adaptability by supporting dynamic, easily customizable cybersecurity scenarios without substantial overhead. Instead of relying on complex, agent-based frameworks that require dedicated infrastructure, we employ Ansible which offers simplicity, agentless operation, human-readable syntax and modular design paving the way for customization and smooth development. VulnerSec supports both virtual machines and containerized environments, maximizing compatibility across diverse infrastructures. Virtual machines are essential for comprehensive system simulation, permitting in-depth interaction with operating system components, while containers offer an efficient method for swift deployment of vulnerable services. This approach fulfills the first functional requirement for scenario customization (FR1).

Consistent with our commitment to encouraging cooperation and transparency, VulnerSec is released under a permissive open-source license, inspiring educators, researchers and practitioners alike to reuse, contribute to and expand the platform. The complete implementation as well as all scenarios and roles is publicly available on Github [18], thus satisfying the second functional requirement for open-source access (FR2). Table II provides an overview of the example vulnerabilities implemented to date, including a concise description and the associated vulnerability type.

Aware of the importance of long-term usability, we specifically target the handling of scenario rot, which happens when outdated vulnerabilities become unusable due to incompatible software versions or broken dependencies. To counteract this, the framework incorporates containerized deployments, which encapsulate legacy software in isolated, reproducible runtime environments, thus satisfying the third functional requirement for handling long-term sustainability (FR3). For this purpose, we use a role that sets up Docker on the target machine. Vulnerable containers can subsequently be launched using a predefined container template with customizable configuration parameters. Nevertheless, software tools deployed today are readily available as images, so this versatile approach further eases the setup, consisting only of specifying an image and the configuration for the vulnerable container. It allows VulnerSec to be compatible and incorporate third-party containers which are already available, such as the container repository published by VulHub. This further demonstrates the flexibility of the proposed framework.

In accordance with the fourth functional requirement for ease of use (FR4), each vulnerability is implemented as an Ansible role, structured according to best practices. These environments are self-contained where each role envelopes a specific vulnerability, ensuring educators and researchers can deploy complex scenarios without navigating tangled dependencies. Deploying a vulnerability is as simple as running the corresponding Ansible role in the virtualized environment. One notable advantage of the VulnerSec framework is that vulnerabilities can be customized to be largely platform-agnostic, leveraging Ansible's modularity and abstraction capabilities. Because each exploit is encapsulated as an independent Ansible role, instructors can swap modules, override variables, or publish new scenarios without touching the surrounding playbooks. [20]. The declarative format permits instructors to easily adapt scenarios to different base operating systems or infrastructures without requiring significant rewrites or modifications. Recent research suggests that Ansible is highly regarded for its ease of use, attributed to its agentless design and straightforward YAML playbooks. Users with minimal prior experience in configuration management tools have reported smoother onboarding and faster initial deployment processes [19]. Its usability is further reflected in its continued adoption within the cybersecurity domain. Chouliaras et al. conducted a systematic survey of ten cyber ranges that were developed in the last decade, analyzing their characteristics based on deployment strategies, configuration management approaches, and automation tools. Their findings indicate that Ansible is among the most widely adopted solutions for automating the deployment and configuration of cybersecurity scenarios, highlighting its prevalence within the field. Using this approach enables instructors and researchers alike to quickly spin up reproducible and isolated virtual machines from a single configuration file, significantly reducing deployment time.

Beyond accelerating the generation and maintenance of vulnerable instances, the platform's design also lays the groundwork for network monitoring and data collection. With the integration of standard monitoring tools, ranging from packet

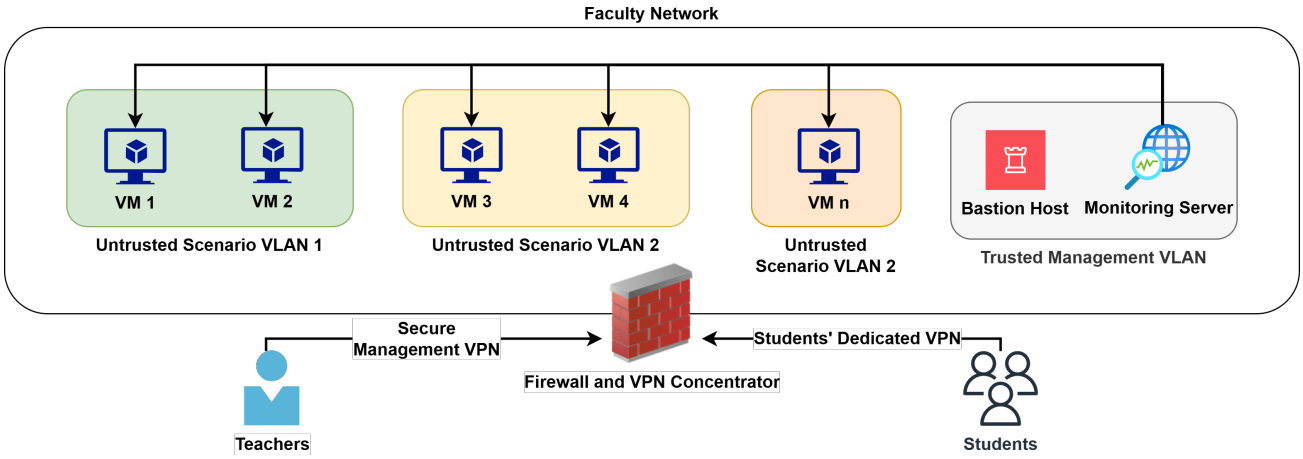


Fig. 1. Visual representation of the network architecture used during the validation stage

sniffers to log analyzers, platform administrators could capture detailed traces of all traffic generated within these isolated environments. This data can then be annotated to identify attempted attacks, the specific vulnerabilities exploited, and whether each intrusion attempt succeeded or failed. Such annotations would offer invaluable insight not only for validating new detection algorithms, but also for building specialized datasets that reflect realistic attack patterns. As a result, the researchers would be able to examine the effectiveness of different defensive techniques under numerous threat scenarios. These integrated monitoring capabilities thus would extend VulnerSec's utility, offering a customizable testbed where both educators and practitioners can structure, assess, and advance security tools and practices in a contained environment.

V. Practical Validation

The proposed VulnerSec framework was practically validated during an introductory course at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, involving 26 students who participated in the hands-on exercises within a one-week timeframe. We initiated the validation process by instantiating 13 virtual machines, each configured with one to three vulnerabilities. The participants' primary objective was to discover the machines and exploit the vulnerabilities. To accommodate remote participation, students conducted the exercises from their homes, connecting securely to the infrastructure via dedicated virtual private network (VPN) profiles. This ensured controlled and isolated interactions with the vulnerable environments while maintaining the integrity and security of the host network.

The infrastructure was hosted on—premise and comprised an OPNSense firewall [21], a VPN server for secure remote student access, an ESXi hypervisor, and Docker container runtimes instantiated in the virtual machines themselves for hosting certain vulnerabilities. All of the virtual machines were continuously monitored to ensure their availability and to verify that the students adhered to the rules and did not make changes to the configuration of the deployed services. While local testing was conducted using VirtualBox, the actual deployment within the faculty's infrastructure utilized the Vagrant provider for VMware. Once instantiated, a bastion host was used to execute the customization playbooks, thus acting as a unified place from where all involved educators could manage the environment.

It should be noted that the provisioning process is decoupled from the deployment process, allowing the same Ansible-based configuration to be applied regardless of the virtualization provider. Fig. 1 provides a visual overview of the described infrastructure setup, illustrating the network topology and undertaken isolation measures.

All of the vulnerabilities were deployed by executing an Ansible playbooks, comprised of dedicated roles corresponding to the scenario at hand. This role-based approach enables flexibility, allowing us to configure different scenarios dynamically. Each virtual machine combined a variety of vulnerabilities, including misconfigured services, outdated web applications and containerized legacy software backdoors. Moreover, a dedicated flag-generation mechanism was implemented, enabling students to verify successful exploitation in a structured manner. Text-based flags are placed in predetermined locations of the file-system, which become accessible to the attacker once the corresponding vulnerability has been successfully exploited. This acts as a further proof that the student has indeed successfully exploited the vulnerability and thus completed the challenge.

To ensure strict isolation from the rest of the network infrastructure, the virtual machines hosting the vulnerabilities were placed in dedicated virtual local area networks (VLANs). The firewall restricted their network access, blocking all inbound and outbound connectivity, with the only exception being the virtual private network subnet. This allowed the students to probe the virtual machines for vulnerabilities, as well as establish reverse connections to their own devices as part of the exploitation process. Multiple VLANs were utilized to

simulate a more elaborate network environment, reminiscent of the real-world.

VI. Conclusion

Cybersecurity education and research requires hands-on practice with realistic attack scenarios, yet existing solutions are often highly complex, non-customized, closed-source, or suffer from scenario degradation. To address these challenges, this paper introduced VulnerSec, an open-source framework for deploying customizable and reproducible virtualized and containerized vulnerable environments. Our contributions include a modular approach that allows flexible scenario creation, a solution to scenario rot by leveraging containerized deployments, and an easily deployable infrastructure, decoupling the provisioning and deployment processes. This framework enables users to efficiently manage cybersecurity environments without extensive overhead.

Future work will primarily focus on expanding the range of vulnerabilities by introducing new exploits as well as developing predefined vulnerability profiles which will allow instructors to generate randomized attack scenarios with different levels of difficulty. Moreover, future possibilities include the generation of events datasets where each captured event is traceable to a specific vulnerability, exploit type, or traffic pattern. Training data of this nature will further contribute to the development of Intrusion Detection Systems (IDS) and other defensive algorithms since a variety of machine learning approaches such as supervised, unsupervised or semi-supervised learning benefit from labeled, authentic scenarios. The flexibility of the framework allows researchers to iterate rapidly, adding or modifying test scenarios to reflect emerging threats and new defense mechanisms. At the same time, it presents the potential for crowdsourcing additional attack traces or analytics drawn from their own testing scenarios. This collective approach would enrich the data beyond what a single group could create on its own, thereby improving the overall quality and research value of the dataset.

As the initial implementation was tested on a limited group of students, future work will also focus on large-scale deployments to evaluate scalability and performance in distributed environments. By releasing VulnerSec under a permissive open-source license, we invite collaboration from the cybersecurity community to refine and extend the platform for the evolving needs of cybersecurity education.

VII. Acknowledgement

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje, under the CYBER-EAGLE project.

This research was in part sponsored by the NATO Science for Peace and Security Programme under grant id. G7593

References

[1] CompTIA, "State of the Tech Workforce 2024," CompTIA, Downers Grove, IL, USA, Mar. 2024. [Online]. Available: https://www.comptia.org/content/research/state-of-the-tech-workforce

[2] C. O. Ekwueme, E. E. Ekon, and D. C. Ezenwa-Nebife, "The Impact of Hands-On-Approach on Student Academic Performance in Basic Science and Mathematics," Higher Education Studies, vol. 5, no. 6, pp. 47-51, 2015. doi:10.5539/hes.v5n6p47.

[3] L. Williams, E. Anthi, Y. Cherdantseva, and A. Javed, "Leveraging Gamification and Game-Based Learning in Cybersecurity Education: Engaging and Inspiring Non-Cyber Students," J. Colloq. Inf. Syst. Secur. Educ., vol. 11, no. 1, Winter 2024.

[4] M. Karjalainen and T. Kokkonen, "Review of Pedagogical Principles of Cyber Security Exercises," Advances in Science, Technology and Engineering Systems Journal, vol. 5, no. 5, pp. 592–600, 2020, doi: 10.25046/aj050572.

[5] S. V. Cole, "Impact of Capture The Flag (CTF)-style vs. Traditional Exercises in an Introductory Computer Security Class," Proceedings of the 27th ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2022), Dublin, Ireland, July 2022, pp. 470–476, doi: 10.1145/3502718.3524806.

[6] [22] A. Cantelli-Forti and M. Colajanni, "Adversarial Fingerprinting of Cyber Attacks Based on Stateful Honeypots," Proc. Int. Conf. Computational Science and Computational Intelligence (CSCI), pp. 19–24, Dec. 2018, doi: 10.1109/CSCI46756.2018.00012.

[7] S. Kara, S. Hizal, and A. Zengin, "Design and implementation of a DEVS-based cyber-attack simulator for cyber security," Int. J. Simul. Model., vol. 21, no. 1, pp. 53–64, 2022, doi: 10.2507/IJSIMM21-1-587.

[8] D. Kim, M. K. Ahn, S. Lee, D. Lee, M. Park, and D. Shin, "Improved cyber defense modeling framework for modeling and simulating the lifecycle of cyber defense activities," IEEE Access, vol. 11, pp. 114187–114206, Oct. 2023, doi: 10.1109/ACCESS.2023.3324901.

[9] E. Russo, M. Ribaudo, A. Orlich, G. Longo, and A. Armando, "Cyber Range and Cyber Defense Exercises: Gamification Meets University Students," Proc. 2nd Int. Workshop on Gamification in Software Development, Verification, and Validation (Gamify '23), San Francisco, CA, USA, Dec. 2023, pp. 1-9, doi: 10.1145/3617553.3617888.

[10] C. Tunc and S. Hariri, "CyberArena: An Open-Source Solution for Scalable Cybersecurity Labs in the Cloud," Journal of Internet Services and Information Security (JISIS), vol. 5, no. 4, pp. 41-59, Nov. 2015.

[11] C. Nelson and Y. Shoshitaishvili, "DOJO: Applied Cybersecurity Education In The Browser," Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024), Portland, OR, USA, March 2024, pp. 930-936. DOI: 10.1145/3626252.3630836.

[12] Z. C. Schreuders, T. Shaw, M. Shan-A-Khuda, G. Ravichandran, and J.Keighley, "Security Scenario Generator (SecGen): A framework for generating randomly vulnerable rich-scenario VMs for learning computer security and hosting CTF events," 1st UK Workshop on Cybersecurity Training & Education (VIBRANT 2015), Liverpool, UK, 2015.

[13] Vulhub, "Vulhub - Pre-Built Vulnerable Docker Environments," [Online]. Available: https://vulhub.org/. [Accessed: 11-Mar-2025].

[14] A. Mills, J. White, and P. Legg, "GoibhniUWE: A Lightweight and Modular Container-Based Cyber Range," Journal of Cybersecurity and Privacy, vol. 4, pp. 615–628, 2024, doi: 10.3390/jcp4030029.

[15] VulnHub, "VulnHub: Providing Vulnerable Virtual Machines for Learning and Practicing Cybersecurity," [Online]. Available: [https://www.vulnhub.com]. [Accessed: 10-Mar-2025].

[16] C. Tunc and S. Hariri, "CLaaS: Cybersecurity Lab as a Service," Journal of Internet Services and Information Security (JISIS), vol. 5, no. 4, pp. 41-59, Nov. 2015.

[17] Cloud Range, "The Leading Cyber Range as a Service," Cloud Range. [Online]. Available: https://www.cloudrangecyber.com/. [Accessed: 12-Mar-2025].

[18] E. Krajchevska, VulnerSec: A Flexible, Automated and Open-Source Cybersecurity Framework, GitHub. [Online]. Available: [https://github.com/ekrajchevska/VulnerSec]. [Accessed: 12-Mar-2025].

[19] A. B. Ojel and J. I. Teleron, "Configuration Management and Automation Tools: A Comparative Analysis and Overview," International Journal of Advanced Research in Arts, Science, Engineering and Management (IJARASEM), vol. 12, no. 1, pp. 174-185, Jan.-Feb. 2025.

[20] N. Chouliaras, G. Kittes, I. Kantzavelou, L. Maglaras, G. Pantziou, and M. A. Ferrag, "Cyber Ranges and TestBeds for Education, Training, and Research," Applied Sciences, vol. 11, no. 4, p. 1809, 2021. [Online]. Available: https://doi.org/10.3390/app11041809

[21] OPNsense® a True Open Source Security Platform and More - OPNsense® Is a True Open Source Firewall and More. Retrieved March 14, 2025, from https://opnsense.org/

Platform for Environmental Data Crowdsensing and Visualization

1st Pance Cavkovski
Faculty of Computer Science and Engineering
University of Ss. Cyril and Methodius
G+D Netcetera
Skopje, North Macedonia
pance.cavkovski@netcetera.com

2nd Slave Temkov
Faculty of Computer Science and Engineering
University of Ss. Cyril and Methodius
Skopje, North Macedonia
slave.temkov@finki.ukim.mk

3rd Aleksandar Nikov
G+D Netcetera
Skopje, North Macedonia
aleksandar.nikov@netcetera.com

4th Petre Lameski
Faculty of Computer Science and Engineering
University of Ss. Cyril and Methodius
Skopje, North Macedonia
petre.lameski@finki.ukim.mk

5th Vladimir Trajkovik
Faculty of Computer Science and Engineering
University of Ss. Cyril and Methodius
Skopje, North Macedonia
vladimir.trajkovik@finki.ukim.mk

***Abstract*—Air pollution remains a critical public health issue, particularly in urban environments where traditional air quality monitoring systems are limited by high costs and sparse spatial coverage. To address these challenges, innovative and scalable monitoring approaches are needed. This paper presents Pulse.eco, a crowd-sensing platform designed to improve air quality monitoring by leveraging a distributed network of low-cost IoT sensors. The platform collects real-time environmental data, including particulate matter (PM1, PM2.5, PM10), temperature, humidity, and noise levels, allowing high-resolution pollution assessment. Pulse.eco uses a microservice-based architecture, LoRaWAN for low-power, long-range communication, and cloud-based analytics for data processing. The platform provides interactive visualizations through a web interface and offers open-access data for researchers, policymakers, and the public. The system has been successfully deployed in multiple urban areas, demonstrating its effectiveness in high-resolution pollution mapping. The real-time data aggregation, anomaly detection, and predictive capabilities of the platform enable better environmental awareness and data-driven decision making.**

***Index Terms*—pollution, air quality, environment, public health, crowd-sensing, data analysis, LoRaWAN, IoT**

I. Introduction

Air pollution in urban areas poses significant health and environmental challenges. It exacerbates respiratory diseases, leading to increased hospitalizations and mortality [1] [2]. Long-term exposure to pollutants is associated with chronic asthma, pulmonary insufficiency, cardiovascular diseases, and cardiovascular mortality [3]. According to the World Health Organization (WHO), air pollution is considered ”the greatest environmental risk to health,” with approximately 4.2 million deaths annually attributed to ambient air pollution [4]. The urban heat island effect, often exacerbated by pollution, further worsens air quality [5]. Air pollution disproportionately affects vulnerable populations, such as children and the elderly, highlighting inequality issues [6]. The primary sources of pollution include transportation, industrial activities, and domestic use of fossil fuels [7].

In general, air pollution remains one of the most serious environmental health risks in the world. Urban areas are especially affected due to high population density, traffic volume, and industry development, which contribute to higher levels of pollutants such as particulate matter (PM), nitrogen dioxide (NO2), and carbon monoxide (CO) [8]. Traditionally, air quality monitoring relies on a network of fixed stations, which are expensive to deploy and maintain. These stations provide high-quality data, but are limited by their sparse spatial distribution, making it challenging to capture the granularity of urban air pollution [9].

A possible and viable approach to overcome these limitations is to introduce a crowd-sensing air pollution platform. By employing a distributed network of low-cost sensors, it can increase the spatial and temporal resolution of the collected environmental data. This data is essential for pollution patterns discovery, public exposure evaluation, and informed decision making for policy and behavior changes [10]. In this paper, we will discuss the Pulse.Eco platform and its services provided for city of Skopje. Since its establishment as SkopjePulse, the platform has grown to operate in multiple countries, providing valuable insights in local air quality conditions and serving as

a model for other urban areas world-wide.

In this paper, we give an overview of the technical aspects that concern the implementation of such a platform. Section 2 presents related work, offering an overview of environmental crowd-sensing platforms implementations. Then we give an overview of the system design and architecture of pulse.eco as the chosen platform example, explaining the different services that are part of it (3). In the next section, we look into the crowd-sensing approach and the different regards to it (4). Then we continue with data monitoring and visualization as part of such platforms, with the different steps of it (5). In the end we give a short conclusion and refer to future work (6).

II. Related work

Crowd-sensing has emerged as a powerful technique for environmental monitoring, offering a scalable and cost-effective alternative to traditional methods. Early works in 2011 laid the foundation for mobile crowd sensing and computing, highlighting its potential for real-time data collection [9]. A few years later the potential of mobile crowd-sensing has been further explored, emphasizing its application in environmental monitoring [11].

The OpenSense project in 2012 introduced a dataset of mobile air quality measurements in Zurich, exploring the possibility of using public transportation systems as platforms for mobile air quality monitoring [12]. This project utilized a combination of static and mobile sensors, showing the benefits of integrating data from different sources for comprehensive air quality monitoring.

In Skopje, North Macedonia, the use static cameras for air quality estimation by monitoring visibility and linking it to air pollution levels was explored [13]. This work highlights an innovative approach to utilizing existing infrastructure (e.g., security and traffic cameras) to gather environmental data, further increasing spatial resolution. The importance of having such a platform in Skopje, its surroundings, and generally in the country is significant [14].

Several projects have successfully implemented crowd-sensing for air quality monitoring. For example the AirSense system uses crowd-sensing to monitor air quality, demonstrating the effectiveness of integrating low-cost sensors with smartphones for data collection [15]. Another great example is the Third-Eye system, which utilizes mobile phones to monitor air quality, taking advantage of the widespread availability of smartphones to improve data collection coverage [16].

The COCAL system introduced a novel approach by integrating low-cost sensors in public transport, creating a dynamic and real-time air quality monitoring network that provided insights which uncovered urban pollution patterns [17]. This system was successful in improving spatial coverage and offering data that is more representative of daily urban conditions.

There are also more recent studies that use this crowd-sensing approach to gather more air quality data [18] [19] [20]. Furthermore, the method has shown to be reliable when it comes to accuracy of the data [21] [22].

Maybe one of the most recognized platform of this kind is sensor.community. It is a global sensor network driven by contributors that create open environmental data. They connect more than 10 thousand sensors distributed mostly in Europe.

The integration of IoT technologies with crowd-sensing platforms has significantly enhanced the capability of these systems to manage and process large volumes of data. Technologies such as LoRaWAN and MQTT are critical for real-time data transmission in IoT environments, ensuring lightweight and efficient communication, which is essential for platforms like pulse.eco [23].

III. System design and architecture

Pulse.eco's architecture is designed to handle the high variability and volume of data gathered by a wide network of sensors. The system is built in a micro-service oriented architecture, which allows each component to be developed, deployed, and scaled independently. Some of the micro-services are also built with internal modular design to provide proper separation of concern while still maintaining necessary performance. This modular approach ensures that the platform remains flexible, resilient, and capable of integrating new sensors, data sources, or analytical tools as needed. (Fig. 1)

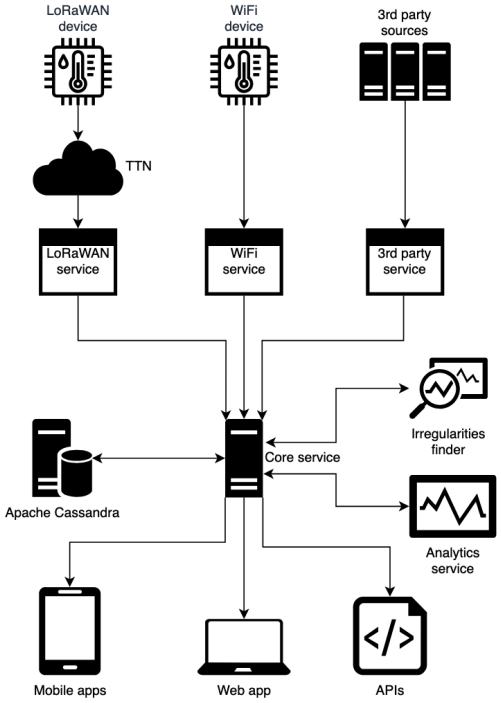


Fig. 1. System overview of Pulse.eco

A. Data Ingestion Services

Couple of services are in charge of data collection from all the registered sensor devices. The data can be transmitted either via WiFi and LoRaWAN protocol, or imported via a 3rd-party provider. LoRaWAN, in particular, is suitable for

long-range communication in both urban and rural settings, providing low-power, wide-area networking capabilities [24]. Each sensor node, whether static or mobile, sends data to gateways, which then relay the data to the cloud for processing and delivery to the final end system. In pulse.eco, TheThingsNetwork community network is used as a LoRaWAN provider.

B. Data Processing and Storage

Data received from the ingestion service undergoes minimal pre-processing, including basic validation, filtering, and transformation. The processed data is then stored in Apache Cassandra, a highly scalable, wide-column, NoSQL database designed to handle large amounts of time-series data efficiently [25].

C. Analytics, irregularities, warnings and Machine Learning

The platform integrates basic statistics, deviations discovery, outage detections along with experimental advanced analytics and machine learning models to derive insights from the collected data. These models include regression analysis, anomaly detection, and predictive analytics using techniques like Long Short-Term Memory (LSTM) networks, which are effective for time-series forecasting [26] [27]. These capabilities allow the platform to predict pollution trends, identify hot-spots, and provide early warnings for increased pollution events, as well as to make sure that the sensors used for input remain active and reliable.

D. Web Application

The default way to preview the information provided by the platform is by using the web application (Fig. 2). pulse.eco provides subdomain access for each individual city registered within it and uses voronoi overlays on top of the city map to visibly display the environmental situation. The maps are provided by OpenStreetMaps with more suitable layers by Stadia maps.

Besides the current situation preview, the web application allows the users to quickly browse all the available environmental data from the past 24 hours. Finally, the Explore section allows them to go back in history and show the situation as it was then.

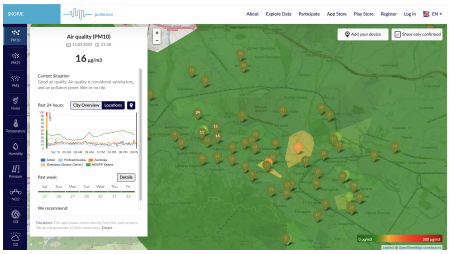


Fig. 2. pulse.eco Web Application

E. API and Data Access Layer

Pulse.eco provides RESTful APIs that allow external applications and users to access the data. This open data approach supports transparency, enabling researchers, policymakers, and the general public to utilize the data for various purposes. In addition to programmatic access, users can execute customized data exports, allowing them to access a broad set of data spanning multiple years.

F. Mobile applications

Not related to the core architecture, but notable components of the overall eco-system, are the client mobile applications. Developed in a native manner for both Android and iOS, the mobile clients use the same API to provide the users with the most needed data for a very seamless and quick access preview.

IV. Crowdsensing approach

The core of Pulse.eco's capability lies in its crowd-sensing approach [28], which involves deploying a large number of low-cost sensors across various locations in a city. These sensors are placed on fixed installations.

A. Participation preconditions

In order to ensure as much reliability of the data, the citizens willing to participate by hosting a sensor device are presented with several conditions prior to the installation:

- the device needs to be placed outdoors (balcony or yard), fixed to a wall or a post, and protected from direct sunlight and rain.
- the network coverage (WiFi or TTN LoRaWAN, based on the type of the device) needs to be solid and reliable
- the device needs to be placed away from active sources of contamination (anything that produces smoke, vibration or sound, e.g.: ashtray, A/C unit, chimney, very close construction site)
- the device needs to face the louder side of the object, so that the noise measurement is more applicable
- the device must not be installed very high (more than 3 or 4 stories)

B. Sensor Specifications

The sensors used in the Pulse.eco network are capable of measuring a range of environmental parameters, including particulate matter (PM1, PM2.5, PM10), temperature, humidity, air pressure and ambient noise levels. Each sensor node consists of properly wired components (Fig. 3) described in the following subsections.

1) ESP32 Microcontroller: A powerful, low-power microcontroller with integrated WiFi and Bluetooth capabilities [29]. It acts as the central processing unit of the sensor node, collecting data from various sensors and managing communication with the backend. In this implementation, the LiLy TTGO development board is used. It comes packed with EP32, LoRa module, antenna connector, OLED mini screen, and multiple exposed GPIO ports as board pins.

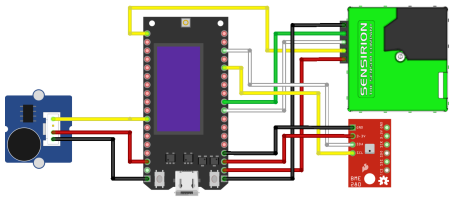


Fig. 3. Wiring diagram of an ESP32 based pulse.eco sensor device

2) Communication Modules: Nodes are equipped with LoRa modules for long-range, low-power communication, making them suitable for wide-area urban monitoring. WiFi is used where a strong local network is available, ensuring reliable data transmission in dense urban settings [23].

3) Particulate matter sensor: Each device is connected to a Sensirion SPS30 PM sensor [30]. The SPS30 is a compact PM10, PM4, PM2.5 and PM1 sensor that operates on laser scattering method and is especially fit for this usage due to his long operating life, self-cleaning capabilities and the ability to power down when not used.

4) Temperature, humidity and air pressure sensor: A composite sensor module, based on the Bosch BME280, is connected to provide data about the ambient temperature, relative humidity, and air pressure [31].

5) Noise sensor: Last, for sensing the ambient noise, a Seeed Sound Sensor is packed in the device configuration as well. The sensor measures raw sound energy which is post-processed to resemble urban noise measure in dBAs [32].

C. Sensor robustness

As the sensor device needs to continuously operate, and it is expected to be placed in rather not that accessible area, it is imperative that it needs to be very robust and to overcome power and network outages, as well as specific sensor element faults. Additionally, the device must not lock itself out of configuration and always be observable. To achieve this, the firmware follows a set of simple state machine transitions (Fig. 4), ensuring that in case of a problem, the device will reboot and keep trying to bring itself into fully operable mode at all times. During device boot, the three separate sensor elements are probed and inspected, and if any of them fails, it's readout and data sending will be omitted. Finally, the firmware presents its own mini web server and an application that is always active. The application uses mDNS to register onto the local network, and presents the user with options to see the overall state, current readouts, operation history, and eventual reconfiguration.

D. Data Collection and Integration

Data is collected continuously from each sensor node, with timestamps and geo-location data, ensuring that each measurement can be accurately mapped. The API allows the devices to report their location, enabling reporting of dynamic, real-time information about pollution levels as these sensor devices can even move through the city, capturing variations in air quality that are not detectable with static sensors alone.

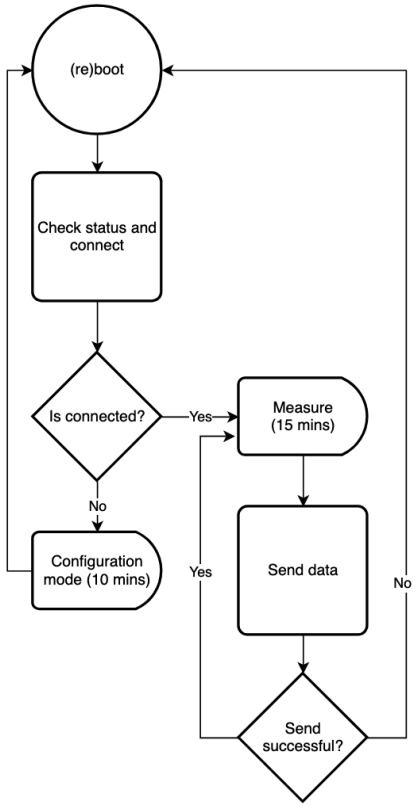


Fig. 4. Sensor device firmware state transitions

E. Data monitoring

In a crowd-sensing scenario, data quality can drift due to various reasons like faulty devices, wrong installation, sensor tainting or miss-use. To combat this, pulse.eco enables automated and manual tools to detect and address outliers. On the basis of a single and multiple days, average and aggregated data from all the devices is compiled. Next, basic statistical products like mean, median and standard deviation are derived, which gives insights on the cumulative state of the devices, isolating the ones that differ from the mass, pointing to possible defects. Ultimately, a single city or a community can have a leader who can take action or contact the device owner to rectify the situation.

V. Application Scenarios

A crowd-sensing platform like pulse.eco provides valuable insights for multiple stakeholders, including city planners, environmental agencies, and the general public. The data collected and analyzed by the platform can be used in various applications, explored in the following subsections.

A. Urban Planning and Policy Making

Detailed air quality maps and pollution forecasts can help city planners design healthier urban environments. This includes optimizing traffic flow, placing green spaces, and regulating industrial activities to minimize pollution exposure. Data-driven decision-making supports the development of policies that reduce emissions and improve overall air quality [33] [34].

Additionally, long-term air quality monitoring can assess the effectiveness of implemented policies, enabling governmental institutions to adapt and continuously improve the urban sustainability. The integration of this kind of data with smart city initiatives can further enhance real-time decision-making, leading to more responsive and efficient environmental management.

B. Public Health Initiatives

Real-time air quality monitoring allows public health officials to issue timely warnings about high pollution levels, advising the public to take preventive measures such as reducing outdoor activities or using masks. Long-term data trends can inform public health strategies aimed at reducing the incidence of pollution-related diseases [10].

Additionally, analysing air quality data combined with health data, professionals can better discover the effect of air pollution on different health conditions. This can help healthcare professionals better allocate their resources and provide personalised health recommendations to vulnerable patients, especially those with pre-existing conditions.

C. Environmental Education and Community Engagement

By making air quality data publicly accessible, Pulse.eco empowers citizens to become active participants in environmental monitoring. Educational programs can use the platform to raise awareness about the impact of air pollution and encourage behaviors that reduce emissions, such as using public transport or cycling [35].

Furthermore, the crowd-sensing approach promotes citizen participation, leading to a more engaged community. This will result in greater public awareness and advocacy for environmental policies, as well as increased collaboration between different people from different profiles. By nurturing transparency and encouraging data-driven discussions, this kind of platform can inspire mass initiatives and community-led efforts to improve air quality.

VI. Conclusion & future work

The Pulse.eco platform is presented in this paper. It demonstrates the effectiveness of a crowd-sensing approach in environmental monitoring, providing a scalable and cost-efficient alternative to traditional air quality monitoring systems. By utilizing a distributed network of low-cost sensors, the platform enables real-time data collection on key environmental parameters such as particulate matter, temperature, humidity, and noise levels. The integration of IoT technologies, including possible LoRaWAN or WLAN communication and cloud-based analytics, ensures reliable data transmission, storage, and processing at scale.

The Pulse.eco platform's modular architecture and open-access data approach enhance its usability for a broad spectrum of stakeholders, including researchers, policymakers, and the general public. The combination of real-time data visualization, historical trend analysis, and predictive capabilities makes Pulse.eco platform a valuable tool for understanding pollution dynamics and supporting data-driven decision-making. Its participatory nature fosters public engagement, encouraging citizens to take an active role in environmental monitoring.

Through its deployment in multiple cities, Pulse.eco has demonstrated its adaptability and potential as a general model for urban air quality monitoring. By bridging the gap between technology and environmental awareness, the platform contributes to improved pollution assessment, informed policy actions, and greater public consciousness about air quality challenges.

All this is crucial for understanding pollution dynamics, evaluating public exposure, and guiding urban planning, public health initiatives and research efforts [10] [12].

Looking forward, expanding the geographical coverage of the sensor network, particularly in under-served regions, is a key priority. Integrating additional sensor types, such as those measuring volatile organic compounds (VOCs) and ozone, will provide a more comprehensive view of air quality. Further improvements in predictive analytics, including the use of machine learning models like Long Short-Term Memory (LSTM) networks, will enable more accurate forecasting of pollution levels, allowing for more timely public health responses and better urban planning [26] [17].

Future work should also focus on improving user engagement through real-time pollution alerts and expanding partnerships with local governments and NGOs to support policy development and environmental education. This kind of platforms also shows a great business value potential which can be explored [36]. Open data initiatives will play a crucial role in empowering researchers and the public to use the platform's data for various applications, contributing to long-term environmental and health studies [9] [13]. As the platform evolves, it has the potential to set a new standard for crowd-sourced environmental monitoring, addressing urban air quality challenges on a global scale.

VII. Acknowledgment

pulse.eco is a project owned and mostly developed by G+D Netcetera. While parts of the system like the devices firmware, hardware design and mobile applications are open-source, the core system is proprietary and built upon G+D Netcetera's IoT framework n-things.

CleanBREATHE is a cooperation of the SPiRIT Research Group at Magdeburg-Stendal University of Applied Sciences (MD-SDL) in Germany and the Faculty of Computer Science and Engineering (FCSE) at the Ss. Cyril and Methodius University in North Macedonia. Many of the new devices

installed in the city of Skopje were sourced as DiY kits, funded by the project. All of these devices were assembled in workshops organized by its members.

REFERENCES

[1] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: A review," *Frontiers in Public Health*, vol. 8, Feb. 2020.

[2] S. Jaganathan, J. D. Bont, M. Dahlquist, Persson, M. Stafoggia, and P. L. S. Ljungman, "Ambient air pollution and cardiovascular diseases: an umbrella review of systematic reviews and meta-analyses," *ISEE Conference Abstracts*, vol. 2021, Aug. 2021.

[3] L. Liang and P. Gong, "Urban and air pollution: a multi-city study of long-term effects of urban landscape patterns on air quality trends," *Scientific Reports*, vol. 10, Oct. 2020.

[4] M. Roser, "Data review: how many people die from air pollution?," *Our World in Data*, 2021. https://ourworldindata.org/data-review-air-pollution-deaths.

[5] R. Cichowicz and A. D. Bochenek, "Assessing the effects of urban heat islands and air pollution on human quality of life," *Anthropocene*, vol. 46, p. 100433, June 2024.

[6] P. Zhou, S. Sun, T. Chen, Y. Pan, W. Xu, and H. Zhang, "Impacts of social inequality, air pollution, rural–urban divides, and insufficient green space on residents' health in china: Insight from chinese general social survey data analysis," *International Journal of Environmental Research and Public Health*, vol. 19, p. 14225, Oct. 2022.

[7] Y. Huang, K. Ropkins, B. Yuan, L. Yao, and L. Huang, "Editorial: Urban air pollution: sources, characterization and mitigation," *Frontiers in Environmental Science*, vol. 11, Dec. 2023.

[8] World Health Organization, "Air quality and health: Fact sheet no. 313," 2011. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health [Accessed: March 2025].

[9] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.

[10] J. J. Henriques, B. Bojarski, K. W. Byrd, and M. Von Wald, "Crowd-sourcing urban air-quality in developing countries through open source technologies," in *2013 IEEE Global Humanitarian Technology Conference (GHTC)*, p. 447–451, IEEE, Oct. 2013.

[11] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, and X. Zhou, "Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm," *ACM Computing Surveys*, vol. 48, p. 1–31, Aug. 2015.

[12] J. J. Li, B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel, "Sensing the air we breathe – the opensense dataset," *AAAI Conference on Artificial Intelligence*, pp. 323–328, 2012.

[13] R. Mingov, P. Lameski, and A. Kulakov, "Pollution estimation based on visibility in urban environments using static cameras: A case study in skopje, macedonia," *International Journal of Environmental Monitoring and Assessment*, vol. 187, p. 41, 2015.

[14] M. Dimovska and R. Mladenovska, "Losing years of human life in heavy polluted cities in macedonia," *Open Access Macedonian Journal of Medical Sciences*, vol. 7, p. 428–434, Feb. 2019.

[15] J. Dutta, F. Gazi, S. Roy, and C. Chowdhury, "Airsense: Opportunistic crowd-sensing based air quality monitoring system for smart city," in *2016 IEEE SENSORS*, IEEE, Oct. 2016.

[16] L. Liu, W. Liu, Y. Zheng, H. Ma, and C. Zhang, "Third-eye: A mobilephone-enabled crowdsensing system for air quality monitoring," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, mar 2018.

[17] P. Diviacco, M. Iurcev, R. J. Carbajales, A. Viola, and N. Potleca, "Design and implementation of a crowdsensing-based air quality monitoring open and fair data infrastructure," *Processes*, vol. 11, no. 7, 2023.

[18] J. Dutta, C. Chowdhury, S. Roy, A. I. Middya, and F. Gazi, "Towards smart city: Sensing air quality in city based on opportunistic crowd-sensing," in *Proceedings of the 18th International Conference on Distributed Computing and Networking*, ICDCN '17, p. 1–6, ACM, Jan. 2017.

[19] H. Elalaoui Elabdallaoui, A. Elfazziki, J. Ourzazi, F. Z. Ennaji, and M. Sadgal, "A crowdsensing-based framework for urban air quality decision support," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, p. 4298–4313, Nov. 2019.

[20] P. Schneider, N. Castell, M. Vogt, F. R. Dauge, W. A. Lahoz, and A. Bartonova, "Mapping urban air quality in near real-time using observations from low-cost sensors and model information," *Environment International*, vol. 106, p. 234–247, Sept. 2017.

[21] C. Zanetti, M. Carraro, M. De Marchi, and S. E. Pappalardo, "Volunteered geographic information for mapping urban climate and air quality: Testing and assessing 'sniffer bikes' with low-cost sensors," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-1/W1-2023, p. 555–561, May 2023.

[22] G. Miskell, J. Salmond, and D. E. Williams, "Low-cost sensors and crowd-sourced data: Observations of siting impacts on a network of air-quality instruments," *Science of The Total Environment*, vol. 575, p. 1119–1129, Jan. 2017.

[23] MQTT, "Mqtt protocol for sensor data communication." Available: https://mqtt.org/ [Accessed: March 2025].

[24] J. Bravo-Arrabal, J. J. Fernandez-Lozano, J. Serón, J. A. Gomez-Ruiz, and A. García-Cerezo, "Development and implementation of a hybrid wireless sensor network of low power and long range for urban environments," *Sensors*, vol. 21, no. 2, 2021.

[25] C. Wang, J. Qiao, X. Huang, S. Song, H. Hou, T. Jiang, L. Rui, J. Wang, and J. Sun, "Apache iotdb: A time series database for iot applications," *Proc. ACM Manag. Data*, vol. 1, jun 2023.

[26] M. Arsov, E. Zdravevski, P. Lameski, R. Corizzo, N. Koteli, K. Mitreski, and V. Trajkovik, "Short-term air pollution forecasting based on environmental factors and deep learning models," *Proceedings of the Federated Conference on Computer Science and Information Systems*, vol. 21, pp. 15–22, 2020.

[27] M. Kostadinov, E. Zdravevski, P. Lameski, P. J. Coelho, B. Stojkoska, M. A. Herzog, and V. Trajkovik, "Forecasting air pollution with deep learning with a focus on impact of urban traffic on pm10 and noise pollution," *PLOS ONE*, vol. 19, p. e0313356, Dec. 2024.

[28] D. E. Boubiche, M. Imran, A. Maqsood, and M. Shoaib, "Mobile crowd sensing – taxonomy, applications, challenges, and solutions," *Computers in Human Behavior*, vol. 101, pp. 352–370, 2019.

[29] Bastelgarage, "Lilygo ttgo lora v1.3 esp32 sx1276 eu868." Available: https://www.bastelgarage.ch/lilygo-ttgo-lora-v1-3-esp32-sx1276-eu868-1-3391 [Accessed: March 2025].

[30] Expo21XX, "Sensirion's sps30 particulate matter sensor awarded mcerts certification." Available: https://www.expo21xx.com/news/sensirions-sps30-particulate-matter-sensor-awarded-mcerts-certification/ [Accessed: March 2025].

[31] B. Sensortec, "Humidity sensors bme280." Available: https://www.bosch-sensortec.com/products/environmental-sensors/humidity-sensors-bme280/ [Accessed: March 2025].

[32] Robocraze, "Sound sensor: Working and its applications." Available: https://robocraze.com/blogs/post/sound-sensor-working-and-its-applications [Accessed: March 2025].

[33] R. Banovic, "Meet the macedonian software engineer fighting air pollution with an app," *Forbes*, 2019. Available: https://www.forbes.com/sites/rebeccabanovic/2019/01/12/meet-the-macedonian-software-engineer-fighting-air-pollution-with-an-app/ [Accessed: March 2025].

[34] E. Mitreska Jovanovska, V. Batz, P. Lameski, E. Zdravevski, M. A. Herzog, and V. Trajkovik, "Methods for urban air pollution measurement and forecasting: Challenges, opportunities, and solutions," *Atmosphere*, vol. 14, p. 1441, Sept. 2023.

[35] V. Batz, P. Lameski, M. A. Herzog, and V. Trajkovik, "Blended research on air pollution using technical and educational solutions - cleanbreathe," *ICT Innovations 2022, Web Proceedings ISSN null*, pp. 108–120, 2022.

[36] F. Brody, M. A. Herzog, V. Trajkovikj, A. Stamboliev, and V. Batz, "Improving air quality as a business: Global markets and local stakeholders," in *Proceedings of the international conference "Economic and Business Trends Shaping the Future"*, EBTSF, Ss Cyril and Methodius University, Faculty of Economics-Skopje, Dec. 2023.

Chronic Kidney Disease Detection Using Mass Spectrometry Data: A Survey

1st Hana Hasanicaj
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
hana.hasanicaj.1@students.finki.ukim.mk

2nd Riste Stojanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
riste.stojanov@finki.ukim.mk

***Abstract*—The human body is very complex, and it is constantly confronted with irregularities or diseases that threaten its quality of life. One particular disease of such nature is Chronic Kidney Disease (CKD), which is characterized by gradual renal failure due to the inability to filter blood properly. Conventional diagnostic tools like serum creatinine-based eGFR and urine albumin-to-creatinine ratio (ACR) have limitations and are not as efficient when detecting early-stage CKD. In order to surpass those limitations, Mass Spectrometry (MS) is being used as a new diagnostic tool. With promising potential in biomarker discovery, it increases the potential for early detection of CKD. Numerous Machine Learning (ML) techniques are used for CKD prediction; however, deep learning networks are used less frequently due to the complexity of the mass spectrometry data. This paper provides an overview of the ML techniques used to process the mass spectrometry data. Moreover, it aims to identify promising ways to use Convolutional Neural Networks (CNNs) in CKD detection.**

***Index Terms*—Chronic kidney disease, mass spectrometry, machine learning, deep learning networks, convolutional neural networks**

I. Introduction

Chronic Kidney Disease (CKD) is a disease that develops progressively; it is irreversible and affects millions of people worldwide. It is characterized by a gradual decline in kidney function, which, if left untreated, can lead to end-stage renal disease (ESRD), requiring dialysis or kidney transplantation. In order to prevent the consequences of the progression of this disease, early detection is crucial. Measuring serum creatinine levels and estimating glomerular filtration rate (eGFR) are commonly used traditional diagnostic methods. However, these methods have limitations when it comes to early detection of CKD, which poses a danger for patients. These traditional biomarkers are not always sensitive or specific enough to differentiate between different stages of kidney dysfunction, highlighting the need for more precise diagnostic approaches [1]–[4]. Mass Spectrometry (MS) is a valuable tool that can be applied in research related to disease detection, including CKD. Its function is to detect and identify biomarkers and molecular signatures with high specificity, sensitivity, and reproducibility [1], [5]. MS-based technologies help researchers discover biomarkers for better early diagnosis and classification of CKD by probing proteins and metabolites from blood, urine, and tissue homogenates [2], [6], [7]. There are new MS tools that have improved the process of CKD biomarker identification, such as the timsTOF Pro equipped with Parallel Accumulation–Serial Fragmentation (PASEF) [5], [8]. MS significantly improves biomarker detection and, thus, impacts disease diagnosis. However, the MS data is quite complex and creates new challenges regarding the data analysis aspect of the problem. Those challenges include high dimensionality, noise, and batch effects [9]–[11]. In order to overcome these obstacles, the data needs to be preprocessed and normalized before it can be used for classification using effective data engineering techniques [9], [12], [13]. Traditional Machine Learning (ML) methods, like Support Vector Machines (SVM) and Random Forests (RF), have proven helpful in classifying MS data to detect diseases [13]–[23]. However, recently, we have been face to face with a very promising alternative for analyzing MS data, and those are deep learning approaches, especially Convolutional Neural Networks (CNNs). CNNs are highly efficient at identifying complex spatial patterns from images that other ML models overlook. According to recent studies, MS data can be transformed into image-like formats, making CNNs the perfect candidates for spectral data analysis [10], [24]–[26]. This paper aims to systematically review the state-of-the-art methods for CKD detection using mass spectrometry data, focusing on both traditional ML approaches and emerging deep learning approaches. The objectives of this survey are to:

- Analyze and assess the range of techniques applied to MS data for CKD detection, highlighting their strengths and limitations
- Provide a comparative analysis of traditional ML versus deep learning models in terms of data preparation, classification accuracy, and clinical applicability
- Locate challenges unique to MS data, such as high dimensionality, noise, and data scarcity, and analyze how these challenges can be addressed
- Outline future research directions for integrating MS data with deep learning techniques to improve CKD detection
- Explore the possibility of using CNNs on MS data by transforming spectral information into structured image-like formats

This comprehensive overview is intended to inform researchers

and clinicians about the current scope of CKD detection using MS data, and the improved diagnosis and patient outcomes as a result of the selection of appropriate analytical methods.

II. Literature Review

A. Mass Spectrometry in Chronic Kidney Disease Detection

1) Data Preprocessing: Preprocessing is necessary for MS data to be usable. Numerous data-engineering techniques, including baseline correction, peak detection, noise reduction, and data normalization, are key steps in MS data preparation. These data-engineering techniques clean the data, enabling signal clarity and feature extraction. This is a necessary step preceding ML analysis [6], [7], [9], [11]. Moreover, some techniques such as signal deconvolution and adaptive filtering are often used to improve signal-to-noise ratio (SNR), enhance reproducibility, and remove batch effects [8], [27], [28]. Whereas data-sampling strategies have been applied to alleviate class imbalance in CKD cohorts [29].

2) Feature Extraction Techniques: Feature extraction transforms high-dimensional spectral signals into lower-dimensional representations while preserving relevant biological information. Some commonly used techniques include Principal Component Analysis (PCA), which reduces tens of thousands of m/z features to fewer components with minimal information loss [4], [12]. Dimensionality reduction techniques like t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) have also been used to maintain local structures in metabolomic data [30], [31]. Additionally, wavelet transforms provide multi-resolution analysis that helps extract transient spectral features with better SNR and temporal localization [8], [32]. These approaches enable better classification and enhance the interpretability of the models.

B. Traditional Machine Learning Approaches

1) Support Vector Machines for Mass Spectrometry Data Analysis: Because of their strong performance and promising results in high-dimensional spaces, Support Vector Machines (SVMs) were among the first ML techniques applied to MS data. SVMs work by finding the optimal hyperplane that separates data into different classes, with the goal of maximizing the margin between the closest points of the two classes. This is quite important when we are dealing with binary classification problems, such as differentiating between CKD-positive and CKD-negative samples. SVMs have an important advantage: they are able to handle non-linear decision boundaries through the use of kernel functions, and this is what allows them to model complex relationships in the data. However, SVMs, especially in the context of MS data, require very careful feature selection and dimensionality reduction. Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE) are Feature extraction techniques often used in conjunction with SVMs to reduce the dimensionality of the data and eliminate irrelevant or redundant features. Despite these preprocessing advantages, SVMs have disadvantages, too. They are prone to overfitting; this is a drawback that is more visible when the number of features is noticeably higher than the number of samples. This makes SVMs less suitable for large-scale datasets with substantial noise [3], [7], [9], [13], [14], [33].

2) Random Forests: An Efficient Alternative: The Random Forest (RF) algorithm emerged as a more efficient alternative to SVMs, particularly in cases where large datasets and high-dimensional data are involved. RF is an ensemble learning method based on decision trees, where multiple decision trees are trained on bootstrapped subsets of the data. The results from all trees are aggregated to produce the final prediction. Since individual trees are less likely to capture noise compared to a single decision tree, this process is effective in reducing overfitting and variance. RF models have several advantages over SVMs. They display greater efficiency and can handle larger datasets with fewer concerns about overfitting. Regarding feature selection or dimensionality reduction, RFs do not require the same level of carefulness, making them more robust in handling the inherent complexity of MS data. However, while RFs improve upon earlier methods in terms of scalability and ease of use, they still have limitations. For example, RF models tend to be less interpretable than SVMs or simpler models. When trying to understand the decision-making process, this can pose critical challenges, particularly in clinical applications where model transparency is critical [4], [15], [28], [34]–[39].

3) Gradient Boosting Machines: Enhanced Performance: The next step in the evolution of ML methods for MS data analysis is the use of Gradient Boosting Machines (GBMs), which offer enhanced performance. GBMs, including algorithms like XGBoost and LightGBM, have gained popularity for their high accuracy and predictive capabilities. Unlike Random Forests, which build trees in parallel, GBMs build trees sequentially, where each new tree corrects the errors made by the previous tree. GBMs are able to capture more subtle data patterns due to the sequential learning process. They have shown remarkable performance when it comes to MS data classification, with studies demonstrating that they often outperform SVMs and RFs in terms of predictive accuracy. The ability of GBMs to minimize error and reduce bias makes them highly suitable for detecting biomarkers related to CKD. However, GBMs have disadvantages, too. When the model complexity increases, we face one such disadvantage: their tendency to overfit when they are not adequately tuned. Furthermore, GBMs can be computationally expensive and time-consuming to train, particularly for large MS datasets with thousands of features. Alternative boosting variants such as AdaBoost have also shown promising potential in CKD classification [13], [25], [26], [40]–[43].

C. Deep Learning Approaches

1) Convolutional Neural Networks: CNNs are specialized neural networks designed to process data in grid-like structures. By applying convolutional filters that capture local patterns in the data, CNNs are efficient in learning hierarchical spatial features. In recent years, deep learning techniques,

especially CNNs, have become increasingly used in MS data analysis and processing. The transformation of MS data from its raw format into image-like data format is one of the key innovations that has greatly contributed to applying CNNs to MS data. The MS data is traditionally one-dimensional data with intensity values at different mass-to-charge ratios. However, this data can be reshaped into two-dimensional matrices. CNNs can be used to process these matrices in order to automatically extract features from them, such as peaks and patterns that are indicative of CKD. In MS data, subtle changes in molecular patterns can often indicate diseases and the ability of CNNs to learn complex data relationships makes them effective for this type of data. Moreover, the need for manual feature extraction is eliminated with CNNs, and this is quite important as it can be labor-intensive and prone to bias. According to studies, in terms of classification accuracy and generalization, CNNs outperform traditional ML models. However, CNNs require large amounts of labeled training data and are computationally expensive, making them less accessible for researchers with limited computational resources [8], [10], [24], [25], [35], [44]–[48].

2) Emerging Trends: Transfer Learning and Multi-Omics Integration: As deep learning techniques continue to advance, new methodologies such as transfer learning and multi-omics integration are being explored to further enhance CKD detection. Transfer learning involves pre-training a model on a large dataset and then fine-tuning it on a smaller, task-specific dataset, which can address the problem of limited labeled MS data. This approach has shown success in fields like computer vision and natural language processing and is now being used in the analysis of MS data for CKD detection. Integrating MS data with other multi-omics data types as well, such as proteomics, genomics, and metabolomics holds significant promise. Multi-omics approaches combine data from different sources to provide a more holistic view of the biological mechanisms underlying CKD. By integrating these diverse data types, researchers can improve the accuracy and robustness of ML models, as they can capture a broader range of disease-related features [12], [30], [31], [31], [49]–[53].

III. Methodology

A. Systematic Review Protocol

A systematic search was conducted following PRISMA 2020 guidelines in order to identify articles related to CKD detection using mass spectrometry data and machine learning techniques [54].

B. Information Sources

The following electronic databases were queried without date restrictions:

- PubMed
- IEEE Xplore
- Google Scholar

C. Search Strategy

The following query was used in PubMed and IEEE Xplore: (”chronic kidney disease” OR CKD) AND (”mass spectrometry” OR ”mass spectrometry data” OR MS OR MS/MS) AND (”machine learning” OR ”deep learning” OR CNN OR ”convolutional neural networks” OR ML OR classification). However, with Google Scholar, to limit the search results to a more reasonable number, the query was the following: (”chronic kidney disease”) AND (”mass spectrometry”) AND (”machine learning” OR ”deep learning” OR ”convolutional neural networks”).

D. Study Selection Process

The selection process included retrieving records, analyzing them, and separating them to filter the relevant sources, including the selected ones in this survey. The workflow is illustrated in the diagram below:

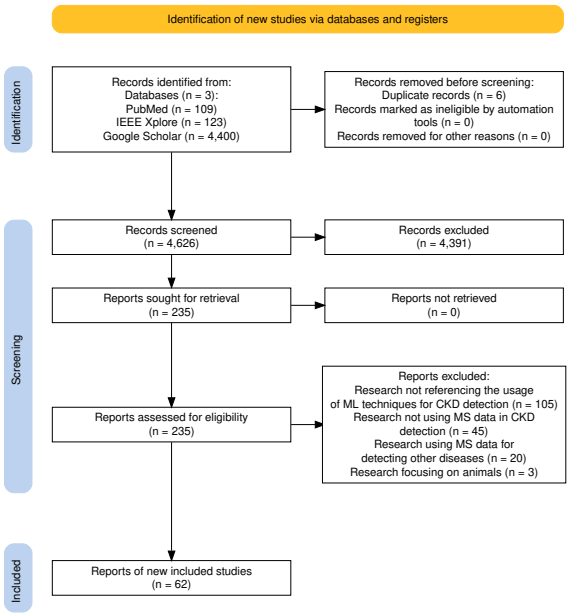


Fig. 1. PRISMA flow diagram for study selection process [55].

IV. Critical Analysis: Convolutional Neural Networks in Mass Spectrometry Data Analysis

CNNs have emerged as powerful tools for MS data analysis, offering distinct advantages over traditional ML methods. Their ability to automatically extract complex hierarchical features from data without the need for manual feature engineering has made them particularly attractive for analyzing high-dimensional datasets like MS spectra. However, while CNNs hold significant promise for improving CKD detection and biomarker discovery, there are both strengths and limitations that need careful consideration [6], [8], [10], [10], [24], [35].

A. Strengths

1) Automated Feature Extraction: Traditional ML models require extensive feature engineering, which is time-consuming and prone to human bias. CNNs, on the other hand, can learn relevant features directly from the raw data, allowing for a more accurate representation of the underlying patterns in MS data. This is particularly beneficial for complex spectral data, where subtle molecular differences can be critical for disease classification [1], [5], [7], [8], [10], [30].

2) Hierarchical Learning: Due to their ability to recognize spatial hierarchies effectively, CNNs can easily identify both basic and complex data patterns. CNNs can uncover intricate molecular details when applied to MS data. It can detect peaks, shifts, and interactions, which traditional models might struggle to detect. In diseases like CKD, where even small molecular changes can be critical, this capacity to capture multi-scale, nuanced features is crucial for identifying biomarkers [2], [10], [44], [46], [47].

3) Improved Accuracy: Studies have shown that CNNs outperform traditional ML algorithms, such as Support Vector Machines (SVMs) and Random Forests (RF) when it comes to classification accuracy for MS data. Their capability to handle large datasets and ability to generalize well across different types of MS spectra make CNNs' particularly effective in detecting CKD-related biomarkers [10], [35], [44], [45], [48], [50], [56].

B. Challenges

1) Data Requirements: The CNNs have a crucial limitation: their requirement for large amounts of labeled data. MS datasets, especially those related to CKD detection, are often limited in size, thus making efficient CNN training challenging. Data scarcity can result in overfitting or poor generalization to unseen data, undermining the model's ability to detect new biomarkers or adapt to different patient populations [13], [31], [34], [57], [58].

2) Computational Demand: CNNs are known for their high computational requirements, particularly when dealing with large and complex MS datasets. The training process can be time-consuming and resource-intensive, requiring specialized hardware such as Graphics Processing Units (GPUs) or cloud-based computing services. Hence, limited computational resources cannot limit the accessibility of CNN-based methods for researchers [4], [10], [31], [47], [49], [59].

3) Interpretability: While CNNs are powerful in extracting features from raw data, they are often described as "black-box" models due to their lack of transparency in making decisions. In biomedical applications such as CKD detection, it is important to understand the reasoning behind a model's predictions to build trust with clinicians and ensure that the model is not making faulty decisions based on spurious patterns in the data. As a result, improving the interpretability of CNN models for MS data analysis is an ongoing challenge [3], [31], [31], [50]–[52].

4) Overfitting: Like many deep learning models, CNNs can be prone to overfitting when trained on small or imbalanced datasets. Even though CNNs are highly capable of detecting complex patterns, if not properly regularized, they may memorize the training data instead of generalizing well to new, unseen data. Techniques like dropout, data augmentation, and transfer learning are essential for mitigating overfitting, but they require careful tuning and optimization [10], [16], [26], [37], [40], [42].

V. Opportunities and Future Directions: Training Convolutional Neural Networks with Image-like Data from Mass Spectrometry

The integration of CNNs into MS data analysis creates new opportunities for improving CKD detection. By transforming spectral data into image-like data formats, new opportunities arise, and CNN can be used quite efficiently to detect hidden molecular patterns in the data [8], [10], [50], [51], [60].

A. Key Opportunities

1) Improved Feature Extraction: Traditional methods rely on manually extracted features, which may miss key molecular patterns. CNNs can automatically learn multi-scale features directly from raw MS data, capturing intricate patterns crucial for detecting early-stage CKD biomarkers [10], [14], [46], [57].

2) Automating Data Analysis: CNNs automate feature extraction, reducing the need for manual intervention. This speeds up the process, enhances reproducibility, and enables high-throughput analysis with transformed MS data [4], [6], [27], [48].

3) Incorporating Advanced Architectures: Future work could enhance CNNs by integrating architectures like recurrent neural networks (RNNs) or attention mechanisms, improving the model's ability to capture long-range dependencies and disease-specific patterns [10], [30], [32], [44], [50].

4) Transfer Learning: Transfer learning, by pre-training CNNs on large datasets and fine-tuning with smaller, disease-specific datasets, could address the challenge of labeled data scarcity, reducing the need for large annotated MS datasets [15], [25], [58], [60].

5) Multi-Omics Integration: Combining MS data with other omics data types (e.g., genomics, proteomics) can enhance disease models by capturing a broader range of disease-relevant features, improving CKD prediction accuracy [2], [3], [40], [49], [61].

B. Future Directions

1) Enhanced Data Augmentation: To improve CNN training with limited MS datasets, data augmentation techniques such as rotation, scaling, and intensity adjustments will be essential. These techniques can expand training data diversity, improve model robustness, and reduce overfitting [12], [37], [42], [62].

2) Real-time Analysis and Diagnostics: CNNs present a transformative opportunity in clinical diagnostics. By optimizing CNNs for real-time MS data analysis, timely disease detection could enable early intervention and personalized treatment strategies, improving patient outcomes [14], [26], [47], [48], [59].

3) Improving Interpretability: Despite their predictive power, CNNs' "black-box" nature hinders interpretability, especially in clinical settings. Future research should focus on methods to visualize and interpret CNN-generated features, ensuring transparency and enhancing clinicians' trust in the model's predictions. By advancing CNNs for image-like MS data transformation, future work could significantly improve CKD detection and lead to personalized healthcare advancements [3], [31], [31], [51], [52], [63].

VI. Real-World Applications of Convolutional Neural Networks in Chronic Kidney Disease Detection

CNNs have emerged as a powerful tool in medical imaging, demonstrating significant potential in detecting and classifying CKD. CNNs' application in CKD diagnosis spans various imaging modalities, including magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound. These CNNs have been shown to successfully classify CKD severity, help in early-stage CKD detection, and identify pathological abnormalities [6], [27], [44], [47], [48], [56]. It is worth mentioning that now, non-invasive methods beyond imaging, such as iris-based CKD screening systems, are also emerging [64]. Despite their advantages and overall success, CNNs are often criticized for their lack of interpretability, commonly called the "black-box" problem. This issue is particularly critical in medical applications, where explainability is crucial for clinical decision-making. In order to address this, researchers have integrated various explainability techniques into CNN-based CKD detection systems. One widely adopted technique is Gradient-weighted Class Activation Mapping (Grad-CAM), which contributes to a model's prediction by generating heatmaps highlighting the most relevant regions in an image. Grad-CAM has been applied to CKD diagnosis to help radiologists and nephrologists understand why a CNN model classifies a kidney as diseased or healthy [25], [31], [40], [51], [52]. Another approach to improving interpretability is SHapley Additive exPlanations (SHAP); this is a technique that assigns importance scores to individual input features. SHAP has been used in CNN-based CKD detection to quantify the contribution of specific biomarkers, pixel intensities, and imaging regions; therefore, as a result, improving model transparency and trust among healthcare professionals. Studies have successfully incorporated SHAP in MS-based CKD detection models, providing insight into which spectral features drive CNN predictions [31], [40]. Moreover, to enhance interpretability and accuracy further, hybrid models combining CNNs with other ML approaches are being explored. Such examples include attention mechanisms and transformer-based architectures. By integrating CNNs with explainability techniques, researchers aim to bridge the gap between high-performance AI models and real-world clinical applications, ensuring that these tools are effective and trustworthy for nephrology practitioners [32], [49], [50], [50], [59], [60], [60].

VII. Conclusion

Applying CNNs to MS data shows potential to be a revolutionary approach to CKD detection in patients. CNNs can automatically learn complex patterns by converting MS spectra into image-like data, leading to more efficient identification of biomarkers and disease-related features. The combination of advanced data augmentation, real-time analysis, and multi-omics integration carries excellent potential for improving diagnostic accuracy and enabling personalized treatments. As these techniques evolve, they promise to revolutionize healthcare, enhancing efficiency and clinical decision-making.

References

[1] M. L. Merchant, "Mass spectrometry in chronic kidney disease research," *Advances in chronic kidney disease*, vol. 17, no. 6, pp. 455–468, 2010, publisher: Elsevier. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1548559510001424

[2] M. Mizdrak, M. Kumrić, T. T. Kurir, and J. Božić, "Emerging biomarkers for early detection of chronic kidney disease," *Journal of personalized medicine*, vol. 12, no. 4, p. 548, 2022, publisher: MDPI. [Online]. Available: https://www.mdpi.com/2075-4426/12/4/548

[3] B. Hocher and J. Adamski, "Metabolomics for clinical use and research in chronic kidney disease," *Nature Reviews Nephrology*, vol. 13, no. 5, pp. 269–284, 2017, publisher: Nature Publishing Group UK London. [Online]. Available: https://www.nature.com/articles/nrneph.2017.30

[4] M. Gayathri Hegde, P. S. Marellavar, G. Sunil Kumar, P. D. Shenoy, K. R. Venugopal, and C. Arvind, "A WebApp Framework for the prediction of e-GFR value and CKD stage using Regression-based Machine Learning Algorithms," in *2024 IEEE 5th India Council International Subsections Conference (INDISCON)*, Aug. 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10744278/

[5] F. Meier, S. Beck, N. Grassl, M. Lubeck, M. A. Park, O. Raether, and M. Mann, "Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device," *Journal of Proteome Research*, vol. 14, no. 12, pp. 5378–5387, Dec. 2015. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jproteome.5b00932

[6] I. Pereira, J. N. S. Sboto, J. L. Robinson, and C. G. Gill, "Paper spray mass spectrometry combined with machine learning as a rapid diagnostic for chronic kidney disease," *The Analyst*, vol. 149, no. 9, pp. 2600–2608, Apr. 2024.

[7] Z. A. Massy, O. Lambert, M. Metzger, M. Sedki, A. Chaubet, B. Breuil, A. Jaafar, I. Tack, T. Nguyen-Khoa, M. Alves, J. Siwy, H. Mischak, F. Verbeke, G. Glorieux, Y.-E. Herpe, J. P. Schanstra, B. Stengel, J. Klein, and CKD-REIN study group, "Machine Learning-Based Urine Peptidome Analysis to Predict and Understand Mechanisms of Progression to Kidney Failure," *Kidney International Reports*, vol. 8, no. 3, pp. 544–555, Mar. 2023.

[8] X. Liu, H. An, W. Cai, and X. Shao, "Deep learning in spectral analysis: Modeling and imaging," *TrAC Trends in Analytical Chemistry*, p. 117612, 2024, publisher: Elsevier. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165993624000943

[9] J. Li, K. Zhou, and B. Mu, "Machine Learning for Mass Spectrometry Data Analysis in Proteomics," *Current Proteomics*, vol. 18, no. 5, pp. 620–634, Nov. 2021. [Online]. Available: https://www.eurekaselect.com/187160/article

[10] Y.-M. Lin, C.-T. Chen, and J.-M. Chang, "MS2CNN: predicting MS/MS spectrum based on protein sequence using deep convolutional neural networks," *BMC Genomics*, vol. 20, no. S9, p. 906, Dec. 2019. [Online]. Available: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6297-6

[11] O. Jardin-Mathé, D. Bonnel, J. Franck, M. Wisztorski, E. Macagno, I. Fournier, and M. Salzet, "MITICS (MALDI Imaging Team Imaging Computing System): a new open source mass spectrometry imaging software," *Journal of proteomics*, vol. 71, no. 3, pp. 332–345, 2008, publisher: Elsevier. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1874391908001127

[12] K. Shankar, P. Manickam, G. Devika, and M. Ilayaraja, "Optimal Feature Selection for Chronic Kidney Disease Classification using Deep Learning Classifier," in *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Dec. 2018, pp. 1–5, iSSN: 2473-943X. [Online]. Available: https://ieeexplore.ieee.org/document/8782340/

[13] C. P. Kashyap, G. S. Dayakar Reddy, and M. Balamurugan, "Prediction of Chronic Disease in Kidneys Using Machine Learning Classifiers," in *2022 1st International Conference on Computational Science and Technology (ICCST)*, Nov. 2022, pp. 562–567. [Online]. Available: https://ieeexplore.ieee.org/document/10040329/

[14] Rajeshwari and H. K. Yogish, "Prediction of Chronic Kidney Disease Using Machine Learning Technique," in *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*, Dec. 2022, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10058678/

[15] N. M. Suganthi, J. V.M, P. Rama, and E. Chandralekha, "Chronic Kidney Disease Detection using AdaBoosting Ensemble Method and K-Fold Cross Validation," in *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Dec. 2022, pp. 979–983. [Online]. Available: https://ieeexplore.ieee.org/document/10029047/

[16] M. U. Devi, M. H. Teja, and S. S. S. R. K, "Chronic Kidney Disease Prediction using Optimized Machine Learning and Ensemble Learning Techniques," in *2024 9th International Conference on Communication and Electronics Systems (ICCES)*, Dec. 2024, pp. 1647–1654. [Online]. Available: https://ieeexplore.ieee.org/document/10859868/

[17] K. M and S. N, "Early Chronic Kidney Disease Detection Using Machine Learning Techniques," in *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*, Mar. 2025, pp. 151–155. [Online]. Available: https://ieeexplore.ieee.org/document/10967970/

[18] D. Baidya, U. Umaima, M. N. Islam, F. M. J. M. Shamrat, A. Pramanik, and M. S. Rahman, "A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2022, pp. 1305–1310. [Online]. Available: https://ieeexplore.ieee.org/document/9776876/

[19] B. Ravindra, N. Sriraam, and M. Geetha, "Chronic Kidney Disease Detection Using Back Propagation Neural Network Classifier," in *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, Feb. 2018, pp. 65–68. [Online]. Available: https://ieeexplore.ieee.org/document/8668110/

[20] G. B. H. Bindu, D. Valluru, H. Vachhani, M. S. Kumar, D. Ganesh, and P. Neelima, "Comparison of Novel Machine Learning Algorithms for Predicting Chronic Renal Disease," in *2024 International Conference on Computational Intelligence for Green and Sustainable Technologies (ICCIGST)*, Jul. 2024, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10717629/

[21] S. Al-Hagree, K. M. Alalayah, M. Alnahdi, M. Alhomdy, A. Alawdi, K. Alhamami, and F. Alsaiari, "An Intelligent System for Diagnosing Kidney Diseases Using Machine Learning," in *2024 1st International Conference on Emerging Technologies for Dependable Internet of Things (ICETI)*, Nov. 2024, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/10777124/

[22] M. Rashed-Al-Mahfuz, A. Haque, A. Azad, S. A. Alyami, J. M. W. Quinn, and M. A. Moni, "Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1–11, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9405681/

[23] F. M. Javed Mehedi Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi, and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Nov. 2020, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/9298026/

[24] N. Bhaskar and M. Suchetha, "An Approach for Analysis and Prediction of CKD using Deep Learning Architecture," in *2019 International Conference on Communication and Electronics Systems (ICCES)*, Jul. 2019, pp. 1660–1664. [Online]. Available: https://ieeexplore.ieee.org/document/9002214/

[25] S. N. Almuayqil, S. Abd El-Ghany, A. A. Abd El-Aziz, and M. Elmogy, "KidneyNet: A Novel CNN-Based Technique for the Automated Diagnosis of Chronic Kidney Diseases from CT Scans," *Electronics*, vol. 13, no. 24, p. 4981, 2024, publisher: MDPI. [Online]. Available: https://www.mdpi.com/2079-9292/13/24/4981

[26] M. Suresh, E. Nirmala, M. Maragatharajan, and M. Yadav, "Chronic Kidney Disease Analysis using Various Biological Parameters under Machine Learning Techniques," in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Dec. 2024, pp. 1449–1455. [Online]. Available: https://ieeexplore.ieee.org/document/10841720/

[27] W. Zhao, Y. Zhang, X. Li, Y. Mao, C. Wu, L. Zhao, F. Liu, J. Zhu, J. Cheng, and H. Yang, "KDClassifier: Urinary Proteomic Spectra Analysis Based on Machine Learning for Classification of Kidney Diseases," *medRxiv*, pp. 2020–12, 2020, publisher: Cold Spring Harbor Laboratory Press. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.12.01.20242198.abstract

[28] V. Shanmugarajeshwari and M. Ilayaraja, "Chronic Kidney Disease for Collaborative Healthcare Data Analytics using Random Forest Classification Algorithms," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Jan. 2021, pp. 1–14, iSSN: 2329-7190. [Online]. Available: https://ieeexplore.ieee.org/document/9402574/

[29] V. D, S. Ramesh, S. K, T. K, S. N, and K. P, "A Machine Learning Perspective for Predicting Chronic Kidney Disease," in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Jul. 2024, pp. 989–993. [Online]. Available: https://ieeexplore.ieee.org/document/10625341/

[30] J. Huang, C. Huth, M. Covic, M. Troll, J. Adam, S. Zukunft, C. Prehn, L. Wang, J. Nano, and M. F. Scheerer, "Machine learning approaches reveal metabolic signatures of incident chronic kidney disease in individuals with prediabetes and type 2 diabetes," *Diabetes*, vol. 69, no. 12, pp. 2756–2765, 2020, publisher: American Diabetes Association. [Online]. Available: https://diabetesjournals.org/diabetes/article-abstract/69/12/2756/16305

[31] K. M. T. Jawad, A. Verma, and F. Amsaad, "Prediction Interpretations of Ensemble Models in Chronic Kidney Disease Using Explainable AI," in *NAECON 2024 - IEEE National Aerospace and Electronics Conference*, Jul. 2024, pp. 391–397, iSSN: 2379-2027. [Online]. Available: https://ieeexplore.ieee.org/document/10670652/

[32] B. Vishnupriya, R. Abitha, A. S. Harshithaa, M. Agalya, and M. Gugapriya, "A Perlustration on the optimistic prognosis of chronic renal failure using Evolutionary Machine Learning and Deep Learning techniques," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Jan. 2023, pp. 1–9, iSSN: 2473-7577. [Online]. Available: https://ieeexplore.ieee.org/document/10128198/

[33] A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, Jul. 2018, pp. 1–9. [Online]. Available: https://ieeexplore.ieee.org/document/8477876/

[34] M. M. S. Raihan, E. Ahmed, A. Karim, S. Azam, M. Raihan, L. Akter, and M. M. Hassan, "Chronic Renal Disease Prediction using Clinical Data and Different Machine Learning Techniques," in *2021 2nd International Informatics and Software Engineering Conference (IISEC)*, Dec. 2021, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9672365/

[35] B. S. Lakshmi, V. Anand R, H. Omotunde, M. Birundadevi, M. S, and M. G. Brahmam, "Deep Learning for Automated Classification of Kidney Lesions in Chronic Kidney Disease Patients," in *2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)*, Feb. 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10498835/

[36] B. P. Saha, S. R. Dash, N. K. Rout, A. P. Patnaik, and M. R. Mishra, "Kidney Disease Prediction using ML techniques," in *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, Feb. 2024, pp. 314–318. [Online]. Available: https://ieeexplore.ieee.org/document/10481648/

[37] R. Bhowmick and A. X. V M, "Machine Learning Models for Analysis and Prediction of Chronic Kidney Disease," in *2023 9th International Conference on Advanced Computing and Communication Systems*

(ICACCS), vol. 1, Mar. 2023, pp. 937–942, iSSN: 2575-7288. [Online]. Available: https://ieeexplore.ieee.org/document/10112872/

[38] B. R.B, R. P, V. K. T, P. P, M. S, and S. K. M, "A Multi Model to Enhance The Detection and Classification of Chronic Kidney Disease Using Machine Learning," in *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, Dec. 2024, pp. 1–9. [Online]. Available: https://ieeexplore.ieee.org/document/10895405/

[39] A. P, A. S. N, B. R, D. C, and E. B, "Data Sampling Methods for Predicting CKD and Diabetes with Machine Learning," in *2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, Jan. 2025, pp. 1–6, iSSN: 2767-9934. [Online]. Available: https://ieeexplore.ieee.org/document/10959341/

[40] M. E. Haque, S. M. J. Islam, J. Maliha, M. S. H. Sumon, R. Sharmin, and S. Rokoni, "Improving Chronic Kidney Disease Detection Efficiency: Fine Tuned CatBoost and Nature-Inspired Algorithms with Explainable AI," in *2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT)*, Mar. 2025, pp. 811–818, iSSN: 2473-5655. [Online]. Available: https://ieeexplore.ieee.org/document/10968421/

[41] M. V N, H. N, J. M, J. S. Reddy, and K. S. K, "Integrating Ensemble and Classical ML Models for Predicting CKD," in *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, Jul. 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10673716/

[42] M. J. Britto, D. Pulugu, R. S. Maragatham, V. Pandey, K. S. Shanthini, and N. P. Sapkal, "The Importance of Early Detection and Risk Assessment for Chronic Disease Based on IF-ORBF-DA Approach," in *2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC)*, May 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10593870/

[43] P. Renuka, C. Thilagavathi, S. Sathiyanathan, L. Murugesh, and M. Madhumitha, "Adaboost Machine Learning Based Modelling to Predict Chronic Kidney Disease Staging," in *2024 International Conference on Science Technology Engineering and Management (ICSTEM)*, Apr. 2024, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/10560884/

[44] M. Revathi, R. Nithiya, S. O. Shankar, and D. Maheshwari, "Prediction Of Chronic Kidney Disease On Ct Images Using Deep Learning: Resnet-34 And Vggnet-16," in *2024 International Conference on Integration of Emerging Technologies for the Digital World (ICIETDW)*, Sep. 2024, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/10939622/

[45] M. S. Hossain, S. M. Nazmul Hassan, M. Al-Amin, M. N. Rahaman, R. Hossain, and M. I. Hossain, "Kidney Disease Detection from CT Images using a customized CNN model and Deep Learning," in *2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS)*, Feb. 2023, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10074314/

[46] A. Balamanikandan, M. Saravanakumar, S. Gunasekaran, V. Anjum, P. Gurusamy, and N. Ashokkumar, "Deep Learning in the Detection of Chronic Kidney Disease," in *2024 4th International Conference on Intelligent Technologies (CONIT)*, Jun. 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10627434/

[47] M. Amanatulla, G. Swathi, M. Pallavi, and K. Bindu, "MRI Scans for Deep Learning-Based Chronic Nephropathy Detection: A Comparison of CNN, MobileNet, VGG16, and ResNet-50 Models," in *2024 5th International Conference for Emerging Technology (INCET)*, May 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10593144/

[48] S. Patil and S. Choudhary, "Prediction of Ultrasound Kidney Imaging Using Convolution Neural Networks," in *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, Apr. 2023, pp. 451–455, iSSN: 2473-5655. [Online]. Available: https://ieeexplore.ieee.org/document/10134748/

[49] P. K. Rao, S. Chatterjee, K. Nagaraju, S. B. Khan, A. Almusharraf, and A. I. Alharbi, "Fusion of Graph and Tabular Deep Learning Models for Predicting Chronic Kidney Disease," *Diagnostics (Basel, Switzerland)*, vol. 13, no. 12, p. 1981, Jun. 2023.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[51] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626. [Online]. Available: http://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

[53] S. S. Vellela, L. R. Vuyyuru, K. B. S. K, N. MalleswaraRaoPurimetla, L. Dalavai, and M. V. Rao, "A Novel Approach to Optimize Prediction Method for Chronic Kidney Disease with the Help of Machine Learning Algorithm," in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, vol. 6, Sep. 2023, pp. 1677–1681. [Online]. Available: https://ieeexplore.ieee.org/document/10397974/

[54] R. Sarkis-Onofre, F. Catalá-López, E. Aromataris, and C. Lockwood, "How to properly use the PRISMA statement," *Systematic Reviews*, vol. 10, no. 1, pp. 117, s13 643–021–01 671–z, 2021. [Online]. Available: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-021-01671-z

[55] N. R. Haddaway, M. J. Page, C. C. Pritchard, and L. A. McGuinness, "PRISMA2020: An r package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis," *Campbell Systematic Reviews*, vol. 18, no. 2, p. e1230, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cl2.1230. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cl2.1230

[56] Y. Zhang and X. Liu, "Machine Learning Techniques for Mass Spectrometry Imaging Data Analysis and Applications," *Bioanalysis*, vol. 10, no. 8, pp. 519–522, Apr. 2018. [Online]. Available: https://www.tandfonline.com/doi/full/10.4155/bio-2017-0281

[57] M. Abirami, S. Das, and R. Sood, "Chronic Kidney Disease Identification Using Random Forest and XGBoost," in *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, May 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10581392/

[58] P. Kiran Rao and S. Chatterjee, "TabNet to Identify Risks in Chronic Kidney Disease Using GAN's Synthetic Data," in *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Oct. 2022, pp. 209–215. [Online]. Available: https://ieeexplore.ieee.org/document/9988284/

[59] A. S. A. Rabby, R. Mamata, M. A. Laboni, Ohidujjaman, and S. Abujar, "Machine Learning Applied to Kidney Disease Prediction: Comparison Study," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2019, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/8944799/

[60] X. Meng, Z. He, L. Guo, H. Lin, and L. Feng, "OSCA-finder: Redefining the assay of kidney disease diagnostic through metabolomics and deep learning," *Talanta*, vol. 264, p. 124745, 2023, publisher: Elsevier. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0039914023004964

[61] M. S. Jayaprabha and V. Vishwa Priya, "Early Prediction of Chronic Kidney Disease: A Comprehensive Survey," in *2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, Jan. 2024, pp. 45–51. [Online]. Available: https://ieeexplore.ieee.org/document/10493896/

[62] T. M. Rahman, S. Siddiqua, S. E. Rabby, N. Hasan, and M. H. Imam, "Early Detection of Kidney Disease Using ECG Signals Through Machine Learning Based Modelling," in *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*, Jan. 2019, pp. 319–323. [Online]. Available: https://ieeexplore.ieee.org/document/8644354/

[63] A. Prasad, V. Asha, B. Saju, S. M. Manoj, P. Manoj Kumar, and B. N. Ramesh, "Machine Learning for Identification of Immedicable Renal Disease," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Jan. 2023, pp. 1–7, iSSN: 2832-3017. [Online]. Available: https://ieeexplore.ieee.org/document/10061031/

[64] S. T, P. M. K, B. S, and M. K. D, "Intelligent Iris based chronic kidney identification system," in *2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, Jan. 2025, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10915256/

Escape to Learn: Assessing the Role of Game-Based Learning in Cybersecurity Education

Maja Videnovik
Center for innovactions and digital education Dig-Ed
Skopje, North Macedonia
maja@dig-ed.org

Vladimir Trajkovik
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, North Macedonia
vladimir.trajkovik@finki.ukim.mk

***Abstract* — This study explores the implementation and effectiveness of an Escape room style educational game designed to enhance cybersecurity awareness among primary school students. The research evaluates student engagement, perceived educational value, and the overall impact of game-based learning on cybersecurity education. The study uses a game-based learning approach, integrating an Escape room style game with cybersecurity-related challenges. The game was implemented with 96 seventh-grade students in North Macedonia, who participated in interactive tasks requiring problem-solving, collaboration, and digital literacy skills. Students' perceptions were collected through a structured questionnaire using a 5-point Likert scale and analyzed for insights into engagement, usability, and learning outcomes. Results indicate that students found the Escape room style game engaging, interactive, and educationally valuable. The approach fostered collaboration, critical thinking, and increased cybersecurity awareness. Despite some challenges related to the complexity of tasks, students expressed high motivation and preference for game-based learning over traditional methods. The findings suggest that incorporating Escape room style games in primary education can effectively enhance cybersecurity literacy.**

Keywords — educational games, cybersecurity, Escape room, game-based learning, primary education

I. INTRODUCTION

In the digital age, cybersecurity education has become essential, particularly for young learners who frequently engage with online platforms. As students increasingly navigate digital spaces for communication, learning, and entertainment, the risks associated with cyber threats, online scams, and privacy breaches have also escalated [1]. Cybersecurity awareness is no longer an optional skill but a necessity, as even primary school students are vulnerable to phishing, social engineering, identity theft, and cyberbullying [2]. However, cybersecurity topics are often perceived as abstract and complex, leading to a lack of interest and engagement in traditional instructional settings. To address this challenge, innovative, interactive, and student-centered learning approaches must be explored.

Game-based learning (GBL) has emerged as a promising pedagogical strategy that enhances student engagement, motivation, and knowledge retention [3] as well as critical thinking skills [4]. Well-designed educational games can create immersive learning environments, enabling students to explore content actively rather than passively receiving information. Among various GBL approaches, Escape room style educational games have gained popularity due to their ability to foster collaboration, critical thinking, and problem-solving skills [5]. Escape rooms require players to work together, decipher clues, and solve challenges in order to progress, making them an ideal tool for inquiry-based learning and real-world problem-solving.

In educational settings, Escape room games have been successfully implemented in subjects such as science, mathematics, and programming [6], [7], [8]. However, their application in cybersecurity education remains underexplored, particularly at the primary school level.

Cybersecurity topics, such as password security, online privacy, and responsible digital behavior, are inherently suited for the challenge-based structure of Escape rooms, where students must demonstrate applied knowledge to advance in the game. By designing cybersecurity-themed Escape rooms, educators can provide students with hands-on experiences, reinforcing key concepts in a memorable and engaging way.

This study explores the design, implementation, and evaluation of an Escape room style educational game aimed at enhancing cybersecurity awareness among primary school students. Despite the increasing use of game-based learning in education, there remains a research gap in its application to cybersecurity education at the primary school level [9].

The game, developed for seventh-grade students in North Macedonia, integrates cybersecurity-related challenges that require players to apply knowledge of online safety principles. The motivation behind this research comes from the need to create more effective learning environments where students can actively engage with cybersecurity content in a realistic and enjoyable manner.

The study is guided by the following research questions:

(1) Engagement and usability: What are students' perceptions of the Escape room game in terms of ease of use, interactivity, and entertainment value?

(2) Educational effectiveness: How does participation in the game influence students' understanding of cybersecurity concepts?

(3) Collaboration and problem-solving: To what extent does the game foster teamwork, logical reasoning, and decision-making skills?

This research contributes to the growing field of game-based learning by providing empirical evidence on the effectiveness of Escape room games in cybersecurity education in primary schools. By integrating playful learning experiences with critical digital literacy skills, this study offers practical insights for educators seeking to improve student engagement and knowledge retention in cybersecurity topics. The findings have broader implications for the adoption of gamification in primary education and the development of interactive learning strategies for complex subjects.

The remaining of this paper is structured as follows: Section 2 describes the methodology, including the game design, implementation process, and data collection approach. Section 3 presents the results, analyzing student engagement, learning effectiveness, and game usability. Section 4 discusses limitations and future work. Finally, Section 5 concludes the study, offering recommendations for future research and potential improvements in game-based cybersecurity education..

II. METHODOLOGY

A. Methodological approach

To assess the effectiveness of a game-based learning approach in fostering an interactive learning environment, we designed and implemented an Escape room style educational game with seventh-grade students from a primary school in North Macedonia. The game focused on cybersecurity education, an essential component of digital literacy that equips students with the knowledge and skills necessary to navigate online spaces safely. Early exposure to cybersecurity concepts helps students develop positive digital habits, promoting safe internet usage and responsible digital citizenship [1]. By integrating core cybersecurity topics, including internet safety, the identification and avoidance of online scams, data protection, and ethical online behavior, students are better prepared to address increasingly complex digital challenges as they grow older [10]. Beyond individual skill development, cybersecurity education fosters a culture of digital responsibility, contributing to long-term awareness and preparedness for an increasingly technology-driven world [11].

The Escape room style game was conducted within a school environment, designed to simulate real-world cybersecurity scenarios through an interactive, challenge-based format. To progress through the game, students had to use mobile phones to scan QR codes placed at different locations within the school building, each of which unlocking a new cybersecurity challenge. The game consisted of ten multiple-choice questions, carefully designed to ensure a balanced level of difficulty across all challenges. Upon completion of the game, students were asked to fill out a structured questionnaire to capture their attitudes, perceptions, and overall experiences with the game-based learning approach. The implemented cybersecurity Escape room style educational game flowchart is visually represented in Fig. 1.

B. Educational elements in the game

The educational focus of the implemented game-based approach was cybersecurity, a topic of increasing relevance as students spend significant time online, engaging in digital communication and information sharing. Given the growing risks of cyber threats, it is crucial to equip students with the necessary skills to navigate the digital world safely. Through this game-based learning experience, students were expected to develop an understanding of positive online behavior, including protecting personal information, respecting digital privacy, and identifying insecure links.

To ensure alignment with the curriculum and learning objectives, the Escape room style game incorporated multiple-choice challenges that addressed key cybersecurity issues. A set of ten carefully designed questions guided students through real-life scenarios, encouraging critical thinking and

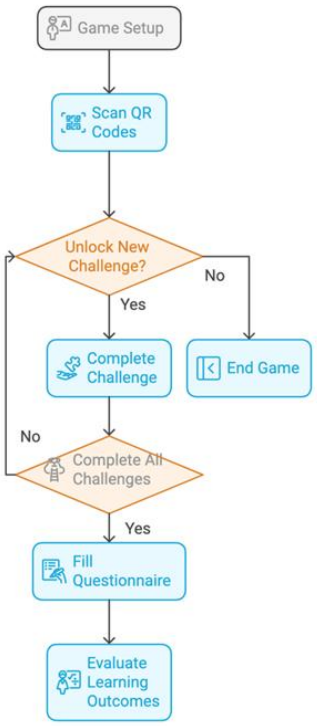


Fig. 1. Cybersecurity Escape room-style educational game flowchart

decision-making. These questions address key topics such as the importance of keeping personal passwords confidential, strategies for protecting data on social networks, and the characteristics of strong passwords. Additionally, the scenarios explore ethical dilemmas and appropriate responses to online risks, including encounters with suspicious messages, phishing attempts, and inappropriate requests from online acquaintances. Emphasis is placed on personal responsibility, respecting others' privacy, and recognizing potentially harmful digital interactions, thereby promoting safe and ethical use of technology among students. The challenges were structured to maintain a balanced level of complexity, ensuring that students of varying digital literacy levels could actively participate and engage with the content.

The primary goal of this research was not to assess students' academic performance or cybersecurity proficiency but to evaluate the effectiveness of game-based learning in fostering engagement and motivation. By analyzing students' attitudes, experiences, and interactions during gameplay, this study provides insights into the potential of interactive learning environments to enhance digital awareness and encourage proactive cybersecurity practices among young learners.

C. Game design

Escape room games have gained popularity as a highly engaging and interactive form of entertainment, renowned for their ability to foster teamwork, leadership, problem-solving, creative thinking, and communication skills [12]. These attributes make them particularly well-suited for educational settings, where they can serve as an effective tool for active learning and student engagement. By integrating Escape room games into the curriculum, educators can create immersive learning experiences that go beyond conventional teaching methods, reinforcing conceptual understanding while simultaneously developing essential 21st-century skills [8].

An educational Escape room is a game-based learning activity in which students work collaboratively to solve a series of challenges within a limited time frame [5]. To progress, participants must identify clues, complete puzzles, and employ strategic thinking to overcome obstacles and achieve a specific goal. When these challenges are directly tied to educational content, students must demonstrate mastery of the subject matter to advance, thereby reinforcing their learning in an engaging and applied context [13]. Incorporating educational elements within the challenges leads to achievement of educational goals and increase in students' motivation and interest in learning [14].

For educators, Escape room games offer a powerful pedagogical tool that transforms traditional lessons into dynamic, student-centered experiences. This method aligns with game-based and collaborative learning principles, providing a structured yet flexible framework that enhances motivation, engagement, and active participation [15]. Escape room style activities have been recognized for their effectiveness in developing problem-solving, lateral thinking, and teamwork skills, as they require students to engage with content through experiential learning and strategic collaboration [16].

To explore the potential of Escape room games in cybersecurity education, a thematic Escape room style game, titled "Cyber Theft", was designed to introduce students to key cybersecurity concepts. The storyline centered around a hacker attempting to steal sensitive data from the school's digital systems. In response, students assumed the role of the school's cybersecurity task force, working collaboratively to solve a series of cybersecurity-related challenges. These challenges assessed students' knowledge of online safety measures, requiring them to apply principles of data protection and digital security to ensure the hacker's attempt was prevented.

The game consisted of five sequential challenges, each composed of multiple-choice cybersecurity questions. To progress through the Escape room, teams had to collaborate, analyze clues, and determine the most appropriate answers. Each successfully completed challenge provided a hint for the next task, encouraging critical thinking and logical deduction.

The Escape room game was structured as a treasure hunt [17], where the correct answers to each challenge revealed the location and password for the next task. To enhance engagement and incorporate mobile technology, students were required to scan QR codes placed at designated locations throughout the school. Each QR code led to a locked digital document, which could only be accessed using a password derived from the correct responses to the previous challenge. This layered problem-solving structure encouraged students to collaborate, apply cybersecurity knowledge, and think strategically to complete the Escape room.

Throughout the game, students were required to demonstrate teamwork, analytical reasoning, and digital literacy as they navigated the challenges. The integration of smartphone-based interactivity not only increased motivation but also provided an authentic real-world application of cybersecurity principles. The Escape room format thus created a highly immersive and educationally valuable experience, reinforcing key cybersecurity concepts in an engaging and interactive manner [18].

D. Evaluation of the effectiveness of the proposed methodological approaches

The game-based learning approach was implemented with 7th grade students at the primary school "Krste Misirkov" in Skopje, North Macedonia, as part of the cybersecurity module. The Escape room style game was conducted within a single class session, with each game lasting 40 minutes. The structured design of the implementation allowed a controlled environment in which students could engage with cybersecurity challenges while ensuring consistency in data collection.

Assessing the effectiveness of game-based learning in an educational context is a multifaceted process that requires an analysis of various parameters. These include student attitudes, engagement levels, interactions during gameplay, and the perceived educational value of the game. To capture these insights, an online survey was conducted immediately after the game session, allowing students to reflect on their experiences while their impressions were still fresh.

The survey design was adapted from previous studies on educational Escape rooms in higher education settings [8] and research on augmented reality games in education [19]. The questionnaire included both quantitative and qualitative measures, enabling a comprehensive evaluation of the game's impact.

At the beginning of the survey, students were asked to provide basic demographic information, including gender, to allow for potential comparative analysis. To quantify their attitudes toward the game, students responded to a series of statements using a 5-point Likert scale, ranging from "strongly disagree" (1) to "strongly agree" (5). The survey focused on key aspects of the game-based learning experience, including:

- Usability and clarity: Whether the game was easy to follow and had clear instructions.
- Interactivity and engagement: Students' perceptions of the game's interactive nature and its ability to capture their interest.
- Challenge and complexity: The perceived difficulty of the questions and the balance between engagement and cognitive effort.
- Collaboration and teamwork: How well the game facilitated group interactions and problem-solving.
- Competitiveness and motivation: Whether students enjoyed the competitive aspect of the game and found it motivating.
- Educational impact: Whether students felt they had learned something new and whether they believed game-based learning was more effective than traditional methods.
- Future applicability: Student interest in using similar educational games in other subjects.

In addition to the Likert-scale responses, students were given the opportunity to provide qualitative feedback through open-ended questions. They were encouraged to share: What aspects of the game they enjoyed the most; Any challenges or difficulties they encountered; and Suggestions for improving the game design and structure. This approach allowed a

subjective evaluation of student engagement and insights into the factors that influenced their learning experience.

By combining quantitative and qualitative data, this evaluation framework provides a holistic understanding of how game-based learning influences student engagement, motivation, and perceived educational value. The results serve as a valuable resource for refining and enhancing future implementations of cybersecurity-focused educational games.

III. Results and Discussion

The methodological approach was implemented with 96 primary school students, and their feedback was gathered through a structured questionnaire assessing their experience with the Escape room style educational game. The responses were analyzed based on student engagement, ease of use, educational value, enjoyment, and future applicability of the game-based learning method.

The results (see Table 1) indicate that students found the Escape room style game highly engaging, with an overwhelming majority expressing enthusiasm for the interactive format. Nearly all students appreciated the fact that the game was played outdoors and on mobile devices, highlighting the novelty and excitement of an active, technology-enhanced learning experience.

More than 97% of students reported enjoying the game, and a large proportion stated that they were impatient to progress through the challenges, underscoring the game's ability to sustain attention and motivation.

While students widely agreed that the game was well-structured and engaging, their responses regarding ease of use were slightly more varied. Approximately 49% of students strongly agreed that the game instructions were easy to follow, while 42% agreed, and 9% remained neutral. The game's sequential challenge structure, requiring students to solve puzzles before advancing, may have contributed to a perceived increase in difficulty, particularly for students with less experience in problem-solving tasks.

TABLE I. Students' attitudes towards implemented Escape room style eductional game

Statement	***Mean***	***SD***
I like that the game is played outdoor	4.89	0.32
I like that the game is played on mobile devices	4.75	0.44
It was easy to follow the instructions of the game. (The game had clear instruction)	4.40	0.66
I like that the game is interactive	4.84	0.36
The game was good organized	4.83	0.37
The questions in the game were not very difficult to solve.	4.22	0.87
I was impatient to open the next question and reach the end.	4.84	0.36
It was fun to play the game	4.88	0.36
I have learned something new with the game.	4.83	0.37
I enjoy that we have cooperated to find the right answer	4.86	0.34
I like competitiveness with other groups	4.82	0.38
I prefer to learn using games rather than a traditional way of learning	4.85	0.35
Using this way of learning through a game, I achieve the desired results faster and more successfully (learn more)	4.79	0.41
I would like to use this kind of game on other classes, too	4.93	0.26
I would recommend this type of game to other students	4.94	0.24

A significant proportion of students expressed confidence in having learned new concepts through the game. Consistent with prior research on educational Escape room games [6], [20], students perceived the problem-solving elements as beneficial for their understanding of cybersecurity topics. However, opinions on the difficulty of the questions were more diverse.

Some students found the questions challenging but rewarding, while others suggested that varying levels of complexity could be incorporated to cater to different skill levels. The standard deviation of responses for question difficulty was higher than other measures, indicating a broader spectrum of opinions on the cognitive demand of the tasks.

One of the most appreciated aspects of the game was its collaborative nature. Students highlighted teamwork and communication as key components of their experience. Additionally, the game fostered a healthy sense of competition, with many students expressing excitement about competing against other groups.

This combination of collaboration and competition tends to increase the level of students engagement [21]. These findings also align with previous studies on the impact of gamification on social interaction [22], demonstrating the potential of Escape room games to promote collaborative problem-solving in educational contexts.

When asked whether they would like to use game-based learning in other subjects, 100% of students either strongly agreed or agreed. The majority expressed a preference for this approach over traditional learning methods, citing higher motivation, engagement, and perceived effectiveness in knowledge retention. Furthermore, many students suggested expanding the number of questions and incorporating different levels of challenge to make future implementations even more engaging.

To ensure data reliability, the questionnaire responses were analyzed using Cronbach's α coefficient, which resulted in a high reliability score ($\alpha = 0.8758$). This confirms that the assessment instrument provided consistent and meaningful insights into student perceptions of the educational Escape room game

IV. Limitations and Future Work

While the findings of this study provide valuable insights, certain limitations should be acknowledged. First, the study was conducted within a single primary school, which may limit the generalizability of the results. Future research could expand the sample size and include students from diverse educational settings to strengthen the findings. The impact of gender on students' attitudes toward the applied game-based learning approach should also be examined.

Second, this study focused primarily on student attitudes and engagement, without directly measuring knowledge retention and skills development over time. Future studies could incorporate longitudinal assessments to evaluate the long-term impact of game-based learning on cybersecurity literacy.

Additionally, further design improvements could be explored to enhance the adaptability of the game. Introducing artificial intelligence based personalized learning paths and incorporating adaptive difficulty mechanisms could provide a more inclusive and differentiated learning experience [23].

V. CONCLUSION

This study explored the implementation and effectiveness of an Escape room style educational game for teaching cybersecurity in primary schools. By integrating interactive, challenge-based learning with real-world cybersecurity scenarios, the game successfully fostered student engagement, collaboration, and critical thinking. The findings suggest that game-based learning can serve as an effective pedagogical tool for enhancing digital literacy and promoting responsible online behavior among young learners.

The evaluation results demonstrate that students found the Escape room game highly engaging, with a strong preference for this method over traditional learning approaches. The integration of mobile technology and outdoor gameplay contributed to increased motivation and active participation. Furthermore, students appreciated the collaborative problem-solving aspects, reinforcing prior research on the benefits of gamification in education. Despite the overall positive response, the findings indicate that some students found the challenge level demanding, suggesting the need for adaptive difficulty levels to accommodate varying skill levels.

From an educational perspective, the study underscores the potential of Escape room style games as an effective approach for cybersecurity education. By embedding learning objectives within an immersive and interactive environment, educators can provide students with hands-on experiences that enhance both cognitive and social skills. The study also highlights the importance of well-structured game design, ensuring that instructional materials are accessible, engaging, and appropriately challenging.

REFERENCES

[1] E. Amankwa, "Relevance of cybersecurity education at pedagogy levels in schools," J. Inf. Secur., vol. 12, no. 4, pp. 233-249, 2021.

[2] Z. Yan, Y. Xue, and Y. Lou, "Risk and protective factors for intuitive and rational judgment of cybersecurity risks in a large sample of K-12 students and teachers," Comput. Hum. Behav., vol. 121, p. 106791, 2021.

[3] Z. Yu, M. Gao, and L. Wang, "The effect of educational games on learning outcomes, student motivation, engagement and satisfaction," J. Educ. Comput. Res., vol. 59, no. 3, pp. 522-546, 2021.

[4] M. H. Hussein, S. H. Ow, L. S. Cheong, and M.-K. Thong, "A Digital Game-Based Learning Method to Improve Students' Critical Thinking Skills in Elementary Science," IEEE Access, vol. 7, pp. 96309–96318, 2019.

[5] M. Wiemker, E. Elumir, and A. Clare, "Escape room games: 'Can you transform an unpleasant situation into a pleasant one?'," in Game Based Learning, J. Haag et al., Eds. St. Pölten, Austria: Fachhochschule St. Pölten GmbH, 2015, pp. 55–68.

[6] S. Y. Huang, Y. H. Kuo, and H. C. Chen, "Applying digital escape rooms infused with science teaching in elementary school: Learning performance, learning motivation, and problem-solving ability," Think. Skills Creat., vol. 37, p. 100681, 2020.

[7] N. Dietrich, "Escape classroom: The LeBlanc process—An educational 'escape game'," J. Chem. Educ., vol. 95, no. 6, pp. 996–999, 2018.

[8] S. López-Pernas, A. Gordillo, E. Barra, and J. Quemada, "Examining the use of an educational escape room for teaching programming in a higher education setting," IEEE Access, vol. 7, pp. 31723-31737, 2019.

[9] M. Videnovik, A. M. Bogdanova, and V. Trajkovik, "Game-based learning approach in computer science in primary education: A systematic review," Entertain. Comput., p. 100616, 2023.

[10] R. L. Suson, "Appropriating digital citizenship in the context of basic education," Int. J. Educ. Learn. Dev., vol. 7, no. 4, pp. 44-66, 2019.

[11] M. D. Richardson, P. A. Lemoine, W. E. Stephens, and R. E. Waller, "Planning for cyber security in schools: The Human Factor," Educ. Plan., vol. 27, no. 2, pp. 23-39, 2020.

[12] C. Jiménez, N. Arís, Á. A. Magreñán Ruiz, and L. Orcos, "Digital escape room, using Genial.ly and a breakout to learn algebra at secondary education level in Spain," Educ. Sci., vol. 10, no. 10, p. 271, 2020

[13] G. Dimova, M. Videnovik, and V. Trajkovik, "Using educational escape room to increase students' engagement in learning computer science," in Proc. 17th Int. Conf. Informatics Inf. Technol. (CIIT), 2020.

[14] S. Nicholson and L. Cable, Unlocking the potential of puzzle-based learning: Designing escape rooms and games for the classroom. 2020, pp. 1–200.

[15] E. Sanchez and M. Plumettaz-Sieber, "Teaching and learning with escape games from debriefing to institutionalization of knowledge," in Proc. Games and Learn. Alliance (GALA), Palermo, Italy, Dec. 2018, pp. 242-253.

[16] P. Williams, "Using escape room-like puzzles to teach undergraduate students effective and efficient group process skills," in Proc. IEEE Integr. STEM Educ. Conf. (ISEC), 2018, pp. 254-257.

[17] K. H. Ng, H. Huang, and C. O'malley, "Treasure codes: augmenting learning from physical museum exhibits through treasure hunting," Pers. Ubiquitous Comput., vol. 22, pp. 739-750, 2018.

[18] N. Komalawardhana, P. Panjaburee, and N. Srisawasdi, "A mobile game-based learning system with personalised conceptual level and mastery learning approach to promoting students' learning perceptions and achievements," Int. J. Mobile Learn. Organ., vol. 15, no. 1, pp. 29-49, 2021

[19] M. Videnovik, T. Vold, L. Kiønig, and V. Trajkovik, "Design thinking methodology for increasing quality of experience of augmented reality educational games," in Proc. 18th Int. Conf. Inf. Technol. Based Higher Educ. Train. (ITHET), 2019, pp. 1-9.

[20] A. Gordillo, D. López-Fernández, S. López-Pernas, and J. Quemada, "Evaluating an educational escape room conducted remotely for teaching software engineering," IEEE Access, vol. 8, pp. 225032-225051, 2020.

[21] E. Vlahu-Gjorgievska, M. Videnovik, and V. Trajkovik, "Computational thinking and coding subject in primary schools: Methodological approach based on alternative cooperative and individual learning cycles," in Proc. IEEE Int. Conf. Teach. Assess. Learn. Eng. (TALE), 2018, pp. 77-83.

[22] M. K. Othman and S. K. Ching, "Gamifying science education: How board games enhances engagement, motivate and develop social interaction, and learning," Educ. Inf. Technol., pp. 1-37, 2024.

[23] N. Rane, S. Choudhary, and J. Rane, "Education 4.0 and 5.0: Integrating Artificial Intelligence (AI) for personalized and adaptive learning," SSRN, 2023.

End-to-End control of a F1Tenth vehicle

1st Atanasko Boris Mitrev
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
and
EXO Service Solutions, Skopje, Macedonia
atanasko.mitrev@exo.mk

2nd Andrea Kulakov
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
andrea.kulakov@finki.ukim.mk

***Abstract*—This paper presents an End-To-End control approach, applied for a RaboRacer autonomous racing vehicle using a deep neural network architecture. The proposed method leverages a 1D Convolutional Autoencoder (CAE) to extract compressed feature representations from high-dimensional LiDAR data combined with vehicle orientation values. The learned latent representation is then processed by a long-short-term memory (LSTM) network, which captures temporal dependencies essential for decision-making in dynamic racing environments. Finally, the Twin Delayed Deep Deterministic Policy Gradient (TD3) reinforcement learning algorithm is employed to generate optimal steering and velocity commands based on the LSTM-processed state representation. CAE is separately pre-trained to enhance feature extraction efficiency, while LSTM and TD3 are jointly trained in a reinforcement learning framework, in order to enable adaptive control. The proposed architecture is evaluated in the F1Tenth Gym ROS simulator, where the vehicle learns to navigate a track autonomously, while optimizing performance metrics such as lap time and collision avoidance. Initial experimental results demonstrate the effectiveness of the end-to-end learning approach in achieving robust and efficient control for autonomous racing.**

***Index Terms*—Autonomous vehicles, RoboRacer, CAE, LSTM, TD3, DNN, sensor embeddings.**

I. Introduction

Autonomous racing presents a unique challenge in robotics and artificial intelligence [1], requiring high-speed decision-making, efficient perception of the environment and precise control. The RoboRacer platform, a 1/10th-scale autonomous racing car, has become a widely used platform for evaluating learning-based control strategies in dynamic racing scenarios. Traditional control approaches, such as modular control architectures and PID controllers, rely on precise vehicle dynamics models and carefully tuned controllers, which may struggle to generalize across different tracks and driving conditions. In contrast, deep learning and reinforcement learning (RL) offer a promising alternative for end-to-end autonomous control, allowing vehicles to learn optimal driving policies directly from sensor data.

This paper proposes an end-to-end deep reinforcement learning architecture for RoboRacing racing, leveraging three key components:

1D Convolutional Autoencoder (CAE) [2]: A feature extraction network that compresses high-dimensional LiDAR scans into a compact latent representation, while preserving critical spatial information. This enables efficient processing of the perception data for downstream learning tasks. Long Short-Term Memory (LSTM) [3] Network: A recurrent neural network designed to capture temporal dependencies in sequential data, allowing the model to retain past observations and improve decision-making in dynamic environments. Twin Delayed Deep Deterministic Policy Gradient (TD3) [4]: A state-of-the-art reinforcement learning algorithm used to learn an optimal control policy for steering and velocity commands, ensuring stable and efficient vehicle control. The CAE is pre-trained separately to enhance its feature extraction capabilities, while the LSTM and TD3 networks are trained jointly. The input to the model consists of LiDAR readings (1080 values) and vehicle orientation data (roll, pitch, yaw), forming a 1083-dimensional input vector. The CAE encodes this data into a 128-dimensional state representation, which serves as an input to the LSTM network. The LSTM network learns temporal dependencies across multiple time steps, before passing its output to the TD3 policy network, which generates the control actions.

LSTM and TD3 reinforcement learning training are conducted in the F1Tenth Gym ROS simulator, where the vehicle learns to navigate a racetrack while optimizing key performance metrics such as lap time and trajectory smoothness. Initial training results demonstrate a promising end-to-end control for efficient autonomous racing, leveraging structured feature extraction and memory-based learning.

The remainder of this paper is structured as follows: Section 2 discusses related work in deep learning-based autonomous control and reinforcement learning for racing. Section 3 details the proposed neural network architecture and training methodology. Section 4 presents the experimental setup and evaluation results. Finally, Section 5 concludes the paper with key findings and future research directions.

II. Related Work

M.M Zarrar et al. [5] in their paper, introduce TinyLidarNet, a lightweight 2D LiDAR-based end-to-end deep learning model for autonomous racing. An F1TENTH vehicle using TinyLidarNet won the 3rd place in the 12th F1TENTH Au-

tonomous Grand Prix competition, demonstrating its competitive performance.

E. Ghignone et al. [6] in their paper present TC-Driver, a Reinforcement Learning (RL) approach for robust control in autonomous racing.

P. Koirala and C. Fleming [7] train offline learning-based algorithms and analyze the cross-track performance of agents, focusing on their zero-shot generalization to unseen scenarios and their ability to adapt to dynamic environmental changes.

III. System Architecture and Training Methodology

This section describes the proposed end-to-end neural network architecture for RoboRacer [8] vehicle control, including the Convolutional Autoencoder (CAE), Long Short-Term Memory (LSTM) network, and Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm. In addition, it details the training methodology used in the F1Tenth Gym ROS simulator environment to learn an optimal driving policy.

A. System Architecture

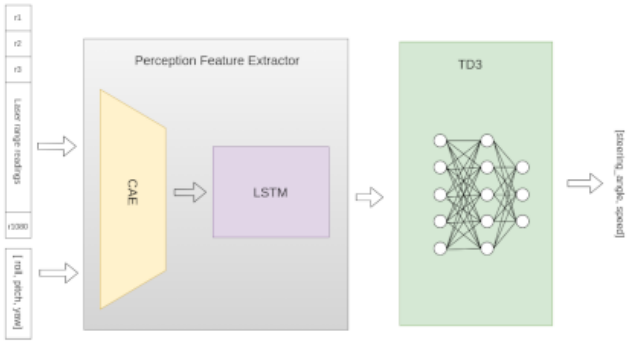


Fig. 1. System architecture

The proposed system processes raw 2D LiDAR sensor data and vehicle orientation (roll, pitch, yaw) to generate optimal steering and speed commands in an end-to-end fashion. The architecture consists of the following components:

1) Convolutional Autoencoder (CAE) for Feature Extraction: The 1D Convolutional Autoencoder (CAE) is responsible for compressing high-dimensional input data into a compact feature representation. The input to the CAE consists of 1083 values (1080 LiDAR range readings + 3 orientation values), which are encoded into a 128-dimensional latent state representation, so-called sensor embeddings. This compression reduces computational complexity while preserving important spatial and structural information from the sensor data.

The encoder consists of convolutional layers with batch normalization and LeakyReLU activation, followed by a fully connected layer:

Conv1D: $1 \rightarrow 64$ (kernel size = 7, stride = 2, padding = 3)
BatchNorm1D 64
LeakyReLU
Conv1D: $64 \rightarrow 128$ (kernel size = 5, stride = 2, padding = 2)
BatchNorm1D 128
LeakyReLU
Flatten
Linear: $128 \times 271 \rightarrow 128$
LeakyReLU

This architecture progressively reduces spatial resolution while increasing the number of feature channels, ensuring that the final latent vector embeddings retain high-level structural information. The CAE is pretrained separately using mean squared error (MSE) loss to ensure effective reconstruction of input data. During the reinforcement learning phase, the weights of the CAE are kept fixed and the encoder is only used to generate the latent state representations of the sensor embeddings, in this phase used for decision-making.

2) LSTM for Temporal Dependency Learning: Since racing involves sequential decision making, a long-short-term memory (LSTM) network is employed to learn the temporal dependencies between the consecutive states. The LSTM processes sequences of consecutive encoded states (from the CAE) and outputs a learned feature vector that represents the temporal context. This allows the network to remember past states, to relate them with the current state, and to make informed control decisions based on the historical information, rather than relying solely on the current LiDAR scan.

3) TD3 for Policy Learning: The Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm is used to train the control policy for the RoboRacer vehicle. TD3 is an actor-critic reinforcement learning algorithm that improves the standard Deep Deterministic Policy Gradient (DDPG) by reducing the overestimation bias and improving stability. The actor network maps the LSTM output to the control actions (steering and speed), while the critic network evaluates the Q-value of the selected actions. TD3 employs delayed policy updates and target noise regularization, making it well-suited for continuous control in high-speed racing scenarios.

B. Training Methodology

The 1D Convolutional Autoencoder (CAE) is trained independently before integrating it into the reinforcement learning pipeline. To generate the training dataset, a PID controller is implemented to autonomously drive the vehicle while following a wall. Data collection is performed on three different maps, with varying starting positions and different reference trajectories to ensure diversity in the dataset. This approach enables the CAE to learn a robust feature representation of the LiDAR data across multiple driving conditions.

The reinforcement learning training is conducted in the F1Tenth Gym ROS simulation environment, where the vehicle learns an optimal driving policy through trial-and-error interactions with the simulated track. The training process consists of the following steps:

1) State Representation:
   - The LiDAR and orientation data are preprocessed and passed through the pretrained CAE encoder to obtain a 128-dimensional latent state sensor embedding.
   - The LSTM network processes a sequence of past sensor embeddings extracted by the CAE, in order to capture the temporal dependencies in the vehicle's perception. The output of the LSTM serves as the state representation for the reinforcement learning algorithm, providing a context-aware encoding that enables a more informed decision-making based on past observations, too.
2) Action Selection:
   - The actor network in the TD3 algorithm takes the LSTM output (the state) as an input and predicts the optimal steering angle and speed commands.
3) Environment Interaction and Reward Computation:
   - The selected action is applied to the F1Tenth vehicle in the simulation, and the next environment state is observed.
   - A reward function is used to guide the learning, encouraging the agent to maximize the speed.
   - The reward is defined as: r = function of(finish the race + do not hit the wall + positive speed + narrow steering)

   Initialize $reward \leftarrow 0$
   If $speed > 0$
   $reward \leftarrow REW_POSITIVE_SPEED$
   If is_crash($lidar_ranges$) **or** $speed < 0$
   $reward \leftarrow reward - REW_CRASH$
   Else
   $reward \leftarrow reward + REW_NO_COLL$
   If is_lap_finished
   $reward \leftarrow reward + REW_FINISH_LAP$
   Return $reward$

4) Policy Optimization:
   - The critic network updates the Q-values using TD3's delayed updates and target smoothing.
   - The actor network is trained to maximize the expected reward.
   - Exploration noise is added during training to encourage policy improvement.
5) Evaluation:
   - The trained policy is evaluated on previously unseen race tracks to test the generalization capabilities of the system.

C. Implementation Details

- Simulation Environment: The training is performed in F1Tenth Gym ROS, using a predefined racetrack.
- Network Architecture: The CAE, LSTM, and TD3 networks are implemented in PyTorch.
- Hyperparameters:
  - CAE: 2 convolutional layers, latent size = 128
  - LSTM: 1 layer, input size = 128, hidden size = 128, output size = 128
  - TD3: Actor-critic learning rate = 0.003, discount factor $\gamma = 0.99$, $\tau = 0.005$, policy noise = 0.2, noise clip = 0.5

IV. Experimental setup and initial results

To evaluate the proposed end-to-end control framework for the RoboRacer vehicle, experiments are conducted in the F1Tenth Gym ROS simulation environment. This section details the training environment, the reinforcement learning setup, the evaluation metrics, and the testing conditions.

A. Simulation Environment

The F1Tenth Gym ROS simulator is used to provide a realistic racing environment with LiDAR-based perception and vehicle dynamics. The vehicle operates in a 2D simulation with 1080 LiDAR range measurements and odometry-based orientation values (roll, pitch, yaw) as input.

Using a PID controller and an implementation of a wall-following algorithm, the data are collected for the purpose of training the CAE. Three different race tracks are selected for collecting the training data to ensure better CAE training and generalization.

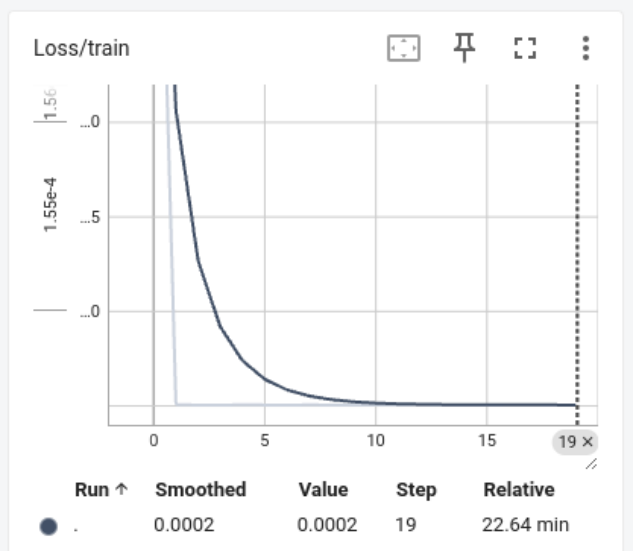


Fig. 2. CAE training results

B. Reinforcement Learning Training

The reinforcement learning model is trained using the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm. The state representation is obtained from pipeline of the 1D Convolutional Autoencoder (CAE) and the LSTM network, capturing both spatial and temporal dependencies in the perception data. The actor network generates continuous control commands:

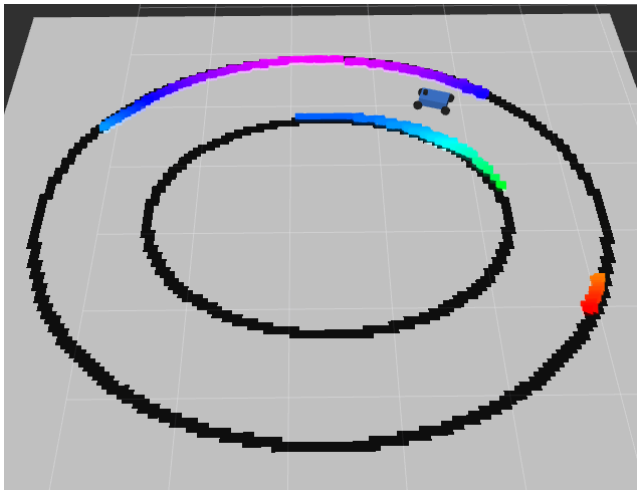


Fig. 3. RL training

- Steering angle $\in [-1, 1]$
- Speed command $\in [-1, 1]$

C. Evaluation Metrics and Testing Conditions

After training, the model is evaluated on unseen track layouts to assess the generalization capability. The key performance metrics include:

- Lap time: Measures the efficiency in completing a lap.
- Track deviation: Evaluates how well the vehicle follows the optimal trajectory.
- Collision rate: Determines the robustness of the learned policy collisions.

V. Conclusion

This paper presents an end-to-end control architecture for autonomous racing vehicle based on a combination of a Convolutional Autoencoder (CAE), Long Short-Term Memory (LSTM) network, and a Reinforcement Learning (RL) algorithm—specifically, Twin Delayed Deep Deterministic Policy Gradient (TD3). The system takes as input a vector of 1083 values, comprising 1080 2D LiDAR readings and 3 orientation values (roll, pitch, yaw). The CAE compresses this high-dimensional input into a compact latent representation, which the LSTM processes to capture temporal dependencies. The resulting sequence embedding is used by the TD3 agent to generate continuous control actions steering angle and speed. The proposed architecture aims to learn a compact spatial representation of the environment while capturing temporal dependencies through sequential modeling, enabling robust autonomous control. Future work will focus on training the model, evaluating its performance in both simulated and real-world environments, and benchmarking it against baseline approaches.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the University Ss. Cyril and Methodius, Skopje, Macedonia and by EXO Service Solutions, Skopje, Macedonia.

References

[1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (International Edition)*. Pearson US Imports & PHIPEs, November 2002. [Online]. Available: http://www.amazon.de/exec/obidos/ASIN/0130803022

[2] D. H. Ballard, "Modular learning in neural networks," in *Proceedings of the Sixth National Conference on Artificial Intelligence*, K. Forbus and H. Shrobe, Eds. San Francisco, CA: Morgan Kaufmann, 1987, pp. 279–284. [Online]. Available: http://www.mpi-sb.mpg.de/services/library/proceedings/contents/aaai87.html

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[5] M. M. Zarrar, Q. Weng, B. Yerjan, A. Soyyigit, and H. Yun, "Tinylidarnet: 2d lidar-based end-to-end deep learning model for f1tenth autonomous racing," 2024. [Online]. Available: https://arxiv.org/abs/2410.07447

[6] E. Ghignone, N. Baumann, and M. Magno, "Tc-driver: A trajectory-conditioned reinforcement learning approach to zero-shot autonomous racing," *IEEE Transactions on Field Robotics*, vol. 1, p. 527–536, 2024. [Online]. Available: http://dx.doi.org/10.1109/TFR.2024.3499912

[7] P. Koirala and C. Fleming, "F1tenth autonomous racing with offline reinforcement learning methods," 2024. [Online]. Available: https://arxiv.org/abs/2408.04198

[8] RoboRacer, "Roboracer," accessed: 28.01.2023. [Online]. Available: https://roboracer.ai/

Evaluating Large Language Models Potential in Mental Health Counseling: A Comparative Study Using the Therapy Empathy Scale

1st Borjan Georgievski
School of Computer Science and Information Technology
University American College Skopje
Skopje, North Macedonia
borjang6@gmail.com

2nd Niki Ramancev
School of Computer Science and Information Technology
University American College Skopje
Skopje, North Macedonia
niki.ramanchev@gmail.com

3rd Marija Stankova Medarovska
School of Computer Science and Information Technology
University American College Skopje)
Skopje, North Macedonia
marija.s.medarovska@uacs.edu.mk

***Abstract*—The increasing demand for accessible mental health support has led to the exploration of digital solutions, including Large Language Models (LLMs), as potential tools for counseling. This study evaluates the capability of an open source LLM from DeepSeek in generating empathetic responses to mental health related queries. Using a dataset of real patient and therapist interactions from online counseling platforms, the study compares LLM generated responses to those provided by certified therapists. The Therapist Empathy Scale (TES) is used to assess the quality and empathy of responses across several key dimensions. Results indicate that DeepSeek's responses consistently scored higher in perceived empathy than those of human therapists across different TES categories. A statistical analysis confirmed significant differences in favor of the LLM generated responses. While these results highlight the potential of LLMs as a supplementary tool in mental health counseling, limitations remain, particularly regarding the absence of human clinical oversight and ethical concerns surrounding AI generated responses. Future research should explore the integration of LLMs into clinical settings, ensuring responsible deployment alongside human therapists.**

Keywords*— *Artificial Intelligence, Large Language Models, Mental Health Support, Empathy Assessment

I. Introduction

Mental health is a growing global concern, with increasing cases highlighting the urgent need for accessible and effective support systems [1]. Psychological counseling with certified therapists provides essential support for individuals facing psychological difficulties. However, significant barriers, such as high costs, limited availability of professionals, and long wait times, often restrict access to traditional therapy, particularly for underserved populations [2]. Given the importance of early intervention and timely diagnosis in promoting mental well-being, addressing these limitations is crucial for improving mental health outcomes.

In response to these challenges, digital mental health services have become more prevalent [3]. Individuals with mobility impairments, irregular work schedules, or financial difficulties are more likely to use digital health services as an alternative to traditional care [4]. Furthermore, people are often more inclined to disclose personal information to artificial intelligence (AI) systems due to reduced fear of judgment [5], and they become more receptive to AI-generated advice as they gain positive experiences with it [6].

Large Language Models (LLMs) have recently demonstrated exceptional capabilities across various tasks, including question-answering and text generation [7]. These models, trained on vast corpora of textual data, generate human-like responses to natural language queries. Their application in mental health care is an emerging field, where their advanced language processing capabilities and extensive knowledge can be leveraged to engage in conversations and offer support to those in need [8]. Several studies have explored the use of LLMs in mental health counseling, indicating their ability to provide meaningful responses and guidance [9, 10, 11].

One of the most critical aspects of effective mental health counseling is therapist empathy, which significantly impacts patient outcomes [12]. Definitions of therapist empathy vary, but they typically highlight the therapist's capacity to understand the patient's experience and communicate this understanding [13]. Previous studies indicate that empathy plays an important role in treatment adherence, increasing patient motivation, improving communication and trust between the patient and therapist, and aligning their views on therapy tasks and goals [12, 13, 14]. Given the importance of empathy in therapy, it is vital to investigate whether LLMs can effectively simulate empathy in a mental health counseling context. A study by Sorin et al. [15] reviewed literature on the capacity of LLMs to demonstrate empathy and reported that, in most cases, LLMs performed comparably to or even surpassed human responses in certain empathy-related tasks. Their conclusion indicated that LLMs could simulate empathetic responses, enhancing user experiences and providing emotionally supportive interactions in various contexts.

A recent study by Ayers et al. [16] examined public forum exchanges where users posted healthcare-related questions to licensed professionals. Their study found that ChatGPT's responses, evaluated by human reviewers, demonstrated greater empathy than those provided by real physicians. Similarly, Lee et al. [17] found that responses from multiple LLM models to common life questions in public forums were consistently rated as more empathetic than human responses by human evaluators. Additionally, Sharma et al. [18] explored how LLM-augmented suggestions impacted perceived empathy, finding that mental health professionals that integrated LLM-generated

suggestions produced responses rated as more empathetic than those without assistance. In another study, Inzlicht et al. [19] concluded that, unlike human professionals, LLMs consistently express empathy without experiencing fatigue and maintain neutrality, offering equal levels of empathy across different groups. Human professionals, in contrast, may exhibit biases, such as preference for their own in-group based on race, religion, or ethnicity.

Given the growing interest in LLMs as potential tools for mental health support, this study aims to assess their ability to provide quality and empathetic responses to patient questions. Specifically, we evaluate the responses from an open-source LLM from DeepSeek [23] and compare them to those of certified therapists, using a dataset of real client-therapist interactions from online counseling platforms [20, 21]. To evaluate the performance, we use the Therapist Empathy Scale (TES) [13], a well-established framework for assessing several dimensions of empathetic communication.

The rest of the paper is structured as follows. Section II discusses the data used in the study, the TES scale for evaluation, and methodology. Section III presents the results in detail; Section IV discusses the findings and Section V provides concluding remarks.

II. Materials and Methods

A. Data

The dataset used for this study is a collection of 3,512 questions and answers sourced from two online platforms for mental health counseling and therapy [20, 21]. The questions cover a wide range of mental health topics, including anxiety, depression, relationship challenges, self-esteem, coping strategies, and more. The responses to the user queries are provided by licensed and qualified mental health professionals, ensuring that the dataset reflects professional therapeutic communication. To adhere to ethical and privacy standards, the data are anonymized, with no personally identifiable information (PII) included.

B. Therapist Empathy Scale (TES)

In mental health counseling, empathy has been identified as an important factor in treatment effectiveness and client outcomes across different psychotherapeutic approaches [13]. Although empathy is often conceptualized as an interactional process between therapist and client, research suggests that it tends to vary more between therapists than between individual therapists and different clients [12]. Therefore, established empathy assessment measures primarily evaluate empathy at the therapist level.

In order to assess the quality and empathy of the LLM generated responses and those of human therapists in addressing patient concerns, the study uses the Therapist Empathy Scale (TES) [13]. TES is a nine-item observer rating scale, which assesses the affective, cognitive, attitudinal, and attunement aspects of therapist empathy. The items are rated on a 7-point Likert scale [22], where lower values indicate absence of the rated component and higher values more extensive presence (1 = not at all, to 7 = extensively). The TES components assess how well therapists understand and respond to clients' feelings and thoughts with acceptance, concern, and openness to change during the counseling session. The TES scale includes the following items:

1. *Concern:* does the therapist convey concern and interest to the client?
2. *Expressiveness:* is the response natural and engaging?
3. *Resonate or capture client feelings:* does the response reflect the client's emotional state?
4. *Warmth:* does the response have a friendly and sincere tone?
5. *Attuned to the client's world:* Does the response demonstrate understanding of the client's emotion?
6. *Understanding cognitive framework:* Does the responses grasp how the client thinks of his problem?
7. *Understanding feelings/inner experience:* does the response acknowledge and validate the client's feeling?
8. *Acceptance of feelings/inner experience:* does the response show non-judgmental acceptance of the client's emotions?
9. *Responsiveness:* does the response address the concern in a relevant way?

C. Large Language Model as a Mental Health Counselor

The study evaluates the ability of an open-source general LLM from DeepSeek, to provide quality and empathetic responses to patient questions related to mental health. While general LLMs gain some exposure during training, they often lack specific training and domain expertise in the field of psychological counseling. More specifically, we evaluate the generated responses from the DeepSeek-R1-Distill-Qwen-14B model, which is a distilled version of the DeepSeek-R1 model [23]. The DeepSeek-R1 model is a 671 billion parameter architecture trained on the DeepSeek-V3-Base LLM using a hybrid supervised fine-tuning (SFT) and reinforcement learning (RL) approach. To enhance usability and reduce computational demands, DeepSeek has also introduced distilled versions of the DeepSeek-R1 model, including DeepSeek-R1-Distill-Qwen-14B, which is a 14 billion parameter model fine-tuned on 800,000 reasoning samples generated by DeepSeek-R1 using the Qwen2.5-14B base. Distilled models (or student models) are smaller, computationally efficient versions of a larger model (or teacher model), created by transferring its knowledge to a reduced parameter architecture [24]. The knowledge distillation process involves training the student model using the teacher model's soft probability outputs, which allows it to approximate the teacher's performance while significantly reducing computational requirements. Fig. 1 provides the prompt used to generate the DeepSeek LLM responses.

D. Evaluation

Evaluating the capabilities of LLMs within specific domains, such as psychological counseling, presents a challenge [9]. Human annotation with domain experts can be costly and time-consuming, making it difficult to scale evaluations efficiently. As a result, researchers have increasingly turned to LLMs as automated evaluators to assess model performance on more open-ended questions [25]. For example, the study of Zheng et al. [26] has found that advanced LLM evaluators match both controlled and crowdsourced human preferences well, at the same level as agreement between humans.

#Prompt

As a professional therapist, provide a warm, thoughtful, and helpful response to the following client concern.

Your response should be:

- Empathetic and validating of the client's feelings.
- Insightful and offering a therapeutic perspective.
- Direct and free of any meta-commentary.
- Focused on both emotional support and practical guidance.
- Personalized to address the specific concerns raised.

Client concern: {context}

Fig. 1. Prompt used for querying the DeepSeek LLM with patient questions related to mental health concerns.

In our evaluation of the quality and perceived empathy in the responses to patient questions from certified human therapists and DeepSeek LLM, we employ Llama LLM from Meta [27]. More specifically, we use the Llama 3.2 version with 3 billion parameters, a distilled version of the larger Llama 3 architecture [28]. For the assessment, Llama serves as an impartial evaluator by providing a rating score from 1 (lowest) to 7 (highest) for each of the nine TES criteria. The model performs a comparative evaluation of responses generated by certified human therapists and the DeepSeek LLM for the same patient query, scoring each aspect of the TES scale. The prompt used for evaluation is presented in Fig. 2. All experiments were conducted on the Google Colab platform using A100 GPUs. To run the DeepSeek-R1-Distill-Qwen-14B and Llama3.2-3B models within the workspace, Ollama was used to pull and deploy the models locally.

III. Results

The evaluation of therapeutic responses generated by certified human therapists and the DeepSeek LLM, using Llama LLM as an automated evaluator across nine categories of the Therapy Empathy Scale (TES), revealed a significant performance difference. As illustrated in Fig. 3, DeepSeek consistently outperformed human therapists in all TES categories, with an average score difference of 2.19 points.

#Prompt

You are a therapeutic expert tasked with evaluating responses based on the Therapist Empathy Scale (TES). Your role is to assess two responses provided to a clients's question. For each response, assign a score from 1 to 7 (1 = not at all, to 7 = extensively) across the following nine TES components:

1. Concern: Does the response demonstrate genuine care and interest in the client's distress?
2. Expressiveness: Is the response delivered in a natural, engaging, and relatable manner?
3. Resonance: Does the response accurately mirror the client's emotional state?
4. Warmth: Does the response convey a friendly, kind, and sincere tone?
5. Attunement: Does the response reflect a deep understanding of the client's inner experience?
6. Cognitive Understanding: Does the response capture the client's thought processes and perception of their problem?
7. Feelings Understanding: Does the response acknowledge and validate the client's emotions?
8. Acceptance: Does the response exhibit non-judgmental acceptance of the client's feelings?
9. Responsiveness: Does the response address the client's concerns in a relevant and meaningful way?

Client Question: {context}
Response1: {response1}
Response2: {response2}

Fig. 2. Prompt used for querying Llama LLM to impartially evaluate responses from human certified therapists and the DeepSeek LLM to patient questions, based on the Therapist Empathy Scale (TES).

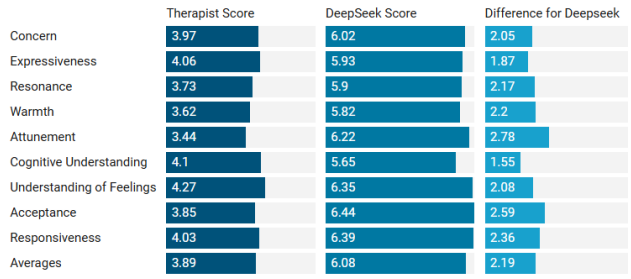


Fig. 3. Comparison of the response quality between (a) certified human therapist and (b) DeepSeek LLM across the nine TES categories using Llama LLM as an automed evaluator.

The mean total TES score across categories for the therapists is 3.89 (SD = 0.82) and for DeepSeek was 6.08 (SD = 0.78), reflecting more perceived empathy in the DeepSeek responses.

To determine the statistical significance of these findings, a Wilcoxon Signed-Rank Test was conducted [29]. This test compared scores of therapist and DeepSeek responses for each of the nine TES categories. Each pair of scores corresponds to the responses on the same patient question by both the therapist and DeepSeek. All p-values were below the significance threshold of $p < .05$, demonstrating statistically significant differences across all categories. This indicates that the observed differences are highly unlikely to be due to random variation.

Further analysis shows that the evaluator preferred the DeepSeek response to the therapist response 93.1% (95% CI, 91.4%-94.8%) of the time. Additionally, the proportion of responses rated as empathetic or very empathetic (scores $\geqslant 4$) was significantly higher for DeepSeek (97.6%, 95% CI, 96.5%-98.7%) compared to therapists (53.0%, 95% CI, 49.5%-56.5%). Analysis of response length showed that DeepSeek's responses (mean = 249.4 words) were considerably longer than those of therapists (mean = 183 words), while patient questions averaged 68.3 words. This suggests that DeepSeek's longer responses may contribute to its higher perceived empathy, potentially by providing more detailed and comprehensive responses. Fig. 4 illustrates an example of dialogue between patient and certified human therapist, as well as between patient and DeepSeek.

IV. Discussion

The results suggest that DeepSeek significantly outperformed certified human therapists in all TES categories based on the evaluation by Llama 3.2. This means that DeepSeek responses were rated as more empathetic, expressive, warm, responsive and understanding than the human therapists in the setting.

While the findings provide valuable insights into the potential of LLMs in therapeutic contexts, it is important to note that the study has several limitations. Firstly, although LLMs are being increasingly used as automated evaluators, the absence of human expert validation necessitates further research to confirm the accuracy and generalizability of these findings. Specifically, human experts in psychotherapy could provide nuanced assessments of empathy that may not be captured by an automated system.

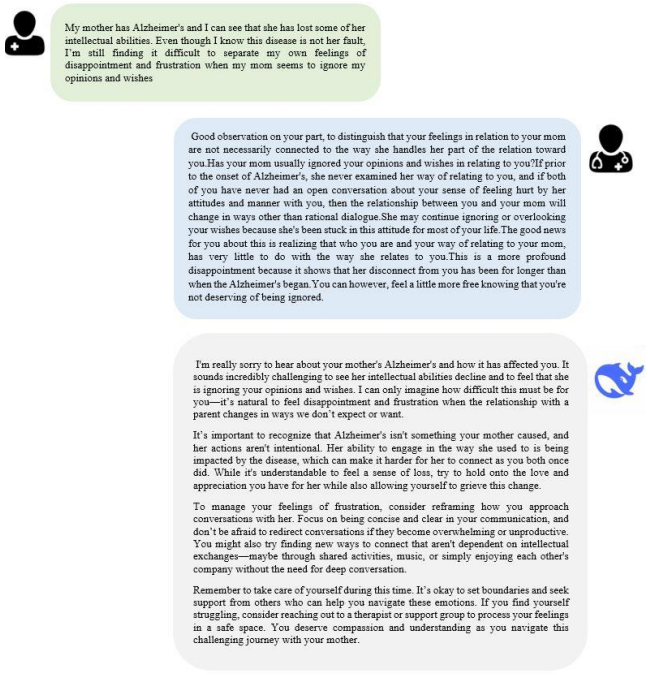


Fig. 4. Comparison of dialogues between: (a) patient and certified human therapist and (b) patient and DeepSeek LLM.

The TES scale, even with a simplified grading system, may not capture all aspects of empathy and therapeutic effectiveness. Secondly, the dataset is sourced from online counseling platforms, which may not fully reflect real-life therapy sessions and limit the generalizability of the results in a clinical setting. Lastly, the usage of LLMs in mental health counseling raises ethical concerns, particularly given their potential to generate fabricated information and lack knowledge when dealing with specific domains, so careful interpretation of their generated responses is needed [30].

V. Conclusion

This study investigated the efficacy of an LLM from DeepSeek in providing empathetic mental health support, by comparing its responses to those of certified human therapists using the Therapy Empathy Scale (TES) and automated evaluation with Llama. The results demonstrate a statistically significant better performance of DeepSeek across all TES categories, suggesting that the LLM's manifest elements of cognitive empathy and provide emotionally supportive responses. This highlights the potential of LLMs to augment mental health support, particularly in therapy access gaps. The ability to generate empathetic responses could improve patient experiences, treatment adherence and foster trust in therapeutic processes. While the findings of the study are promising, further research, particularly by medical experts and within clinical settings, is necessary to better assess their impact on patient outcomes.

References

[1] Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M.R. and Rahman, A., 2007. No health without mental health. The lancet, 370(9590), pp.859-877.

[2] Olfson, M., 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. Health Affairs, 35(6), pp.983-990.

[3] Li, A., Ma, L., Mei, Y., He, H., Zhang, S., Qiu, H. and Lan, Z., 2023. Understanding client reactions in online mental health counseling. arXiv preprint arXiv:2306.15334.

[4] Yeung, A.W.K., Torkamani, A., Butte, A.J., Glicksberg, B.S., Schuller, B., Rodriguez, B., Ting, D.S., Bates, D., Schaden, E., Peng, H. and Willschke, H., 2023. The promise of digital healthcare technologies. Frontiers in Public Health, 11, p.1196596.

[5] Lucas, G.M., Gratch, J., King, A. and Morency, L.P., 2014. It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior, 37, pp.94-100.

[6] Böhm, R., Jörling, M., Reiter, L. and Fuchs, C., 2023. People devalue generative AI's competence but not its advice in addressing societal and personal challenges. Communications Psychology, 1(1), p.32.

[7] Bubeck, S., Chadrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S. and Nori, H., 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712

[8] Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., Sheu, Y.H., Zhou, P., Moran, L.V., Ananiadou, S., Beam, A. and Torous, J., 2024. Large language models in mental health care: a scoping review. arXiv preprint arXiv:2401.02984.

[9] Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y. and Wang, Z., 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. arXiv preprint arXiv:2307.11991.

[10] Liu, J.M., Li, D., Cao, H., Ren, T., Liao, Z. and Wu, J., 2023. Chatcounselor: A large language models for mental health support. arXiv preprint arXiv:2309.15461.

[11] Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A.K. and Wang, D., 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(1), pp.1-32.

[12] Elliott, R., Bohart, A.C., Watson, J.C. and Greenberg, L.S., 2011. Empathy. Psychotherapy, 48(1), p.43.

[13] Decker, S.E., Nich, C., Carroll, K.M. and Martino, S., 2014. Development of the therapist empathy scale. Behavioural and cognitive psychotherapy, 42(3), pp.339-354.

[14] Dhingra, S., Singh, M., Vaisakh, S.B., Malviya, N. and Gill, S.S., 2023. Mind meets machine: Unravelling GPT-4's cognitive psychology. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(3), p.100139.

[15] Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G. and Klang, E., 2024. Large Language Models and Empathy: Systematic Review. Journal of Medical Internet Research, 26, p.e52597.

[16] Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., Faix, D.J., Goodman, A.M., Longhurst, C.A., Hogarth, M. and Smith, D.M., 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine, 183(6), pp.589-596.

[17] Lee, Y.K., Suh, J., Zhan, H., Li, J.J. and Ong, D.C., 2024. Large language models produce responses perceived to be empathic. arXiv preprint arXiv:2403.18148.

[18] Sharma, A., Rushton, K., Lin, I.W., Wadden, D., Lucas, K.G., Miner, A.S., Nguyen, T. and Althoff, T., 2023. Cognitive reframing of negative thoughts through human-language model interaction. arXiv preprint arXiv:2305.02466.

[19] Inzlicht, M., Cameron, C.D., D'Cruz, J. and Bloom, P., 2024. In praise of empathic AI. Trends in Cognitive Sciences, 28(2), pp.89-91.

[20] Bertagnolli, N., 2020. Counsel chat: Bootstrapping high-quality therapy data. Towards Data Science.

[21] Amod. Mental health counseling conversations [Data set]. Retrieved from https://huggingface.co/datasets/Amod/mental_health_counseling_conversations

[22] Joshi, A., Kale, S., Chandel, S. and Pal, D.K., 2015. Likert scale: Explored and explained. British journal of applied science & technology, 7(4), p.396.

[23] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. and Zhang, X., 2025. Deepseek-r1:

Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

[24] Hinton, G., Vinyals, O. and Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[25] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H. and Wang, S., 2024. A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594.

[26] Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36, pp.46595-46623.

[27] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[28] Meta. (2025). Llama 3.2-1B [Model]. Retrieved from https://huggingface.co/meta-llama/Llama-3.2-1B

[29] Meek, G.E., Ozgur, C. and Dunning, K., 2007. Comparison of the t vs. Wilcoxon signed-rank test for Likert scale data and small samples. Journal of modern applied statistical methods, 6(1), p.10.Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., Sheu, Y.H., Zhou, P., Moran, L.V., Ananiadou, S., Beam, A. and Torous, J., 2024.

[30] Zhang, T., Qiu, L., Guo, Q., Deng, C., Zhang, Y., Zhang, Z., Zhou, C., Wang, X. and Fu, L., 2023. Enhancing uncertainty-based hallucination detection with stronger focus. arXiv preprint arXiv:2311.13230.

Hyperparameter Optimization and AutoML of a Simple CNN for Skin Lesion Classification

Marino Jakimoski
Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, Macedonia
marino.jakimoski@students.finki.ukim.mk

Ilinka Ivanoska
Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, Macedonia
ilinka.ivanoska@finki.ukim.mk

***Abstract*—Skin cancer is among the most prevalent cancers worldwide, and its early detection is critical for improving patient outcomes. Convolutional Neural Networks (CNNs) have proven effective in analyzing dermatoscopic images for automated skin lesion classification. However, the performance of CNN models, particularly shallow or simple architectures, heavily relies on hyperparameter selection, which can be a time-consuming and challenging process. This paper investigates the impact of various Hyperparameter Optimization (HPO) techniques and Automated Machine Learning (AutoML) frameworks on a simple CNN model for skin lesion classification. Using the HAM10000 dataset, we evaluated optimization methods such as Bayesian Optimization, Random Search, and SMAC3. Performance metrics, including accuracy, F1 score, and multiclass ROC AUC, were used to assess the models. Our findings demonstrate that while a simple CNN can benefit from HPO, the relatively shallow nature of the network may limit its ability to achieve high performance compared to deeper models. Bayesian Optimization achieved the highest F1 score improvement of 3.9%, but other HPO methods, such as SMAC3, showed less promise for this type of network. These results suggest that more advanced optimization methods or deeper CNN architectures might be more suitable for tasks like skin lesion classification. This work underscores the importance of careful hyperparameter tuning and the potential of AutoML to enhance the performance of simpler CNN models, while also highlighting the challenges faced by shallow networks in complex classification tasks.**

***Index Terms*—Skin lesion classification, Convolutional Neural Networks, Hyperparameter optimization, AutoML, Bayesian optimization, Random search, SMAC3, HAM10000 dataset, Skin cancer detection, Deep learning.**

I. Introduction

Skin cancer is one of the most common cancers worldwide [1]. Detecting it early can save lives, but accurate diagnosis often requires expert analysis of skin images. In recent years, deep learning, particularly Convolutional Neural Networks (CNNs), has emerged as a powerful tool for automatically identifying skin lesions from images. These models can analyze patterns in skin images to assist doctors in making faster and more accurate diagnoses.

However, CNN performance depends heavily on hyperparameters such as learning rate, number and type of layers, and filter sizes. Finding the right combination of these settings can be a challenging and time-consuming process, especially when done manually. Hyperparameter optimization (HPO) methods, including random search and Bayesian optimization, aim to automate this process, leading to more efficient and effective model training. AutoML (Automated Machine Learning) tools take this one step further by integrating hyperparameter tuning with model selection and feature optimization, reducing the need for deep machine learning expertise.

AutoML techniques, particularly HPO, have been widely applied in medical image analysis to automate model tuning. A recent review by Ali et al. [2] highlights optimization strategies like Bayesian methods that improve medical imaging model performance.

Recent advances in deep learning have improved skin lesion classification, with networks like FixCaps [3], MFSNet [4], and attention-based models [5] achieving high performance through advanced architectures and tuning [6]. Class imbalance in datasets like HAM10000 can bias predictions, but HPO and data augmentation have mitigated this issue [7]. Metaheuristic optimization has also enhanced CNN performance, as shown by Golnoori et al. [8], inspiring our focus on HPO.

Despite these advances, simple CNNs, which are suitable for resource-constrained settings, often underperform due to suboptimal hyperparameter settings. The effectiveness of HPO techniques for such models remains underexplored. In this study, we investigate how different HPO techniques can maximize the performance of a simple CNN for classifying skin lesions on the HAM10000 dataset, aiming to identify effective optimization strategies and assess the limitations of shallow architectures.

II. Materials and Methods

The experiments conducted in this research aim to explore the effectiveness of various Hyperparameter Optimization (HPO) techniques and AutoML algorithms to optimize a convolutional neural network (CNN) in the classification of skin lesions. Specifically, we investigate how different optimization methods impact the model's performance in terms of accuracy and other relevant classification metrics. This section outlines the experimental setup, including details about the data, precision settings, and computational constraints.

A. Data

The data used in these experiments is the full HAM10000 dataset [9], which consists of dermatoscopic images representing various skin lesions. The dataset includes a total of 10,015 images that are classified into 7 different categories, representing different types of skin lesions such as melanoma,

basal cell carcinoma, and others. These images are used for both training and testing purposes.

To improve model generalization and avoid overfitting, data augmentation techniques were applied to the training dataset. The augmentations included random scaling, rotation, flipping, and other transformations. These augmentations were applied only to the training images to ensure that the testing set remains unchanged, providing an unbiased and accurate evaluation of the model's performance.

The testing data was used solely for evaluating model performance and was kept separate from the training data to avoid data leakage. This split between training and testing data ensures the validity and reliability of the experimental results.

B. Algorithms

The experiments in this research utilize several Hyperparameter Optimization (HPO) techniques to optimize the performance of a convolutional neural network (CNN) for skin lesion classification. The key algorithms used are as follows:

- **Bayesian Optimization**: Bayesian optimization is a probabilistic model-based method for global optimization, often used for hyperparameter tuning [10]. It builds a surrogate model (e.g., Gaussian process) to approximate the objective function and selects hyperparameters based on an acquisition function to balance exploration and exploitation. In this study, we used the implementation provided by the Keras-Tuner library [11].
- **Random Search**: Random search is a simple and effective method for hyperparameter optimization, where hyperparameters are sampled uniformly at random from a predefined search space [12]. This approach was implemented using the Optuna framework [13], which efficiently handles trial management and search space definitions.
- **SMAC (Sequential Model-based Algorithm Configuration)**: SMAC is a sequential model-based optimization algorithm that uses a Random Forest(RF) as a surrogate model to approximate the objective function [14]. It is well-suited for discrete and continuous hyperparameter spaces. The experiments in this study used the SMAC3 implementation of the algorithm [15].
- Due to hardware limitations, we restricted our study to the above HPO methods. More computationally intensive techniques, such as evolutionary algorithms and swarm optimizations, were attempted but not explored.

Each of these algorithms is used to optimize key hyperparameters of the CNN, including the learning rate, the sizes of the filters, the activation functions, and others, to achieve the best possible model performance. The effect of each algorithm on the CNN's performance is evaluated across various metrics, including accuracy, precision, recall, and F1 score.

C. Baseline CNN Architecture

The baseline CNN model is designed for classifying skin lesions into seven categories with an input shape of $128 \times 128 \times 3$. Table I summarizes the architecture and compilation details.

TABLE I
BASELINE CNN ARCHITECTURE FOR SKIN LESION CLASSIFICATION

Layer Type	**Parameters**
Input	Shape: $128 \times 128 \times 3$
Conv2D	128 filters, 3×3 kernel, ReLU
MaxPooling2D	2×2 pool size
Conv2D	64 filters, 3×3 kernel, ReLU
MaxPooling2D	2×2 pool size
Conv2D	32 filters, 3×3 kernel, ReLU
MaxPooling2D	2×2 pool size
Flatten	-
Dense	128 units, ReLU
Dropout	Rate: 0.5
Dense	7 units, Softmax
Compilation	
Optimizer	Adam, learning rate: 0.001
Loss	Categorical Crossentropy
Metric	Accuracy

D. Trials and Model Evaluation

After setting up the baseline model, we proceeded to perform hyperparameter optimization using all the algorithms separately. Each optimization algorithm was run for 100 trials. For each trial, the following process was followed:

- **Hyperparameter Search and Model Selection**: In each trial, the optimization algorithm searched the hyperparameter space to find the best combination of hyperparameters, including the learning rate, kernel size, and layer width. The best set of hyperparameters was chosen based on the validation performance during the optimization process. For algorithms like Keras-Tuner, the best model was saved for further use.
- **Retraining with Optimal Hyperparameters**: Once the optimal hyperparameters were identified for each trial, the CNN model was reset and trained using the chosen hyperparameters for 20 epochs, just like the baseline model, to ensure a fair comparison in terms of training duration.
- **Model Evaluation**: After retraining, the models were evaluated using three key performance metrics: **accuracy**, **F1 score**, and **multiclass ROC AUC**. These metrics provided a comprehensive assessment of the models' effectiveness in classifying skin lesions and differentiating among multiple classes.

For each optimization algorithm, the best model from all trials was selected based on the highest performance in terms of accuracy, F1 score, and multiclass ROC AUC.

The performance of these optimized models was then compared to the baseline model to determine the effectiveness of hyperparameter optimization, which will be discussed in Section III

III. RESULTS AND DISCUSSION

A. Experimental Setup

To ensure reproducibility, a fixed random seed is used for all experiments. This seed was derived from a Unix timestamp

at the time the first experiment was conducted. Additional setup parameters include using 16-bit floating-point precision (FP16) and setting a graphics memory limit of 7000 MiB. A fixed batch size of 64 was used across all experiments due to memory constraints, and as a result, the batch size was not optimized by the Hyperparameter Optimization (HPO) algorithms. The data was split 90/10 into training and validation. There was no need to create a testing split as the dataset provides a separate set of images for testing.

B. Results

This section presents the results of the experiments described in Section II. We compare the performance of the baseline convolutional neural network (CNN) with models optimized using different hyperparameter optimization (HPO) algorithms. Performance metrics used for evaluation include accuracy, F1 score, and multiclass ROC AUC.

TABLE II
PERFORMANCE METRICS ACROSS OPTIMIZATION METHODS

Optimization Method	Accuracy (%)	F1 Score	AUC
Baseline	67.1	0.618	0.877
Bayesian Optimization	67.4	0.642	0.866
Random Search	67.3	0.630	0.870
SMAC3	4.9	0.051	0.515

TABLE III
KEY HYPERPARAMETER CHANGES ACROSS OPTIMIZATION METHODS

Optimization Method	Key Hyperparameter Changes
Baseline	-
Bayesian Optimization	Filters: [32, 96, 64] Dense Units: 256 Learning Rate: 0.00068
Random Search	Filters: [256, 64, 48] Dense Units: 128 Activation: Tanh
SMAC3	Ineffective Optimization Configuration

The results of hyperparameter optimization (HPO) indicate minor improvements for Bayesian Optimization and Random Search compared to the baseline model, with accuracy increases of 0.3% and 0.2%, respectively (see Table II). Bayesian Optimization achieved the highest improvement, with a 3.9% increase in F1 score (from 0.618 to 0.642), indicating a better balance between precision and recall. Random Search yielded similar but slightly lower improvements (F1: 0.630). Both methods achieved comparable AUC scores, with Bayesian Optimization showing a slight decline over the baseline in OVR AUC, likely due to trade-offs in hyperparameter adjustments. The ROC curves for these methods are shown in Figures 1, 2, and 3, which visually compare their performance. In contrast, SMAC3 drastically underperformed, with an accuracy of only 4.9% and an AUC of 0.515 (see Table II). This poor performance is likely due to the Random Forest surrogate model's inability to effectively model the complex relationships in the hyperparameter space for the CNN architecture and HAM10000 dataset, as illustrated by Figure 4.

The limited impact of HPO across all methods suggests that the shallow CNN architecture may constrain performance improvements. Bayesian Optimization succeeded to a degree because its probabilistic approach efficiently balances exploration and exploitation, identifying hyperparameters that better suit the model's capacity. Random Search, while simpler, also found viable configurations but was less consistent. SMAC3's failure may stem from its reliance on a Random Forest surrogate, which struggled to capture the non-linear interactions in the hyperparameter space for this task. The consistent accuracy plateau across methods indicates that significant performance gains may require deeper architectures or more advanced optimization techniques, as discussed in Section IV.

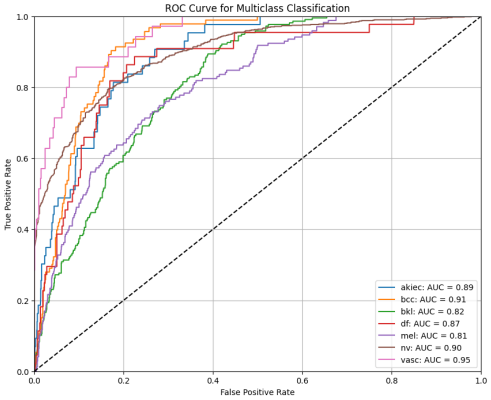


Fig. 1. ROC curve for the baseline model.

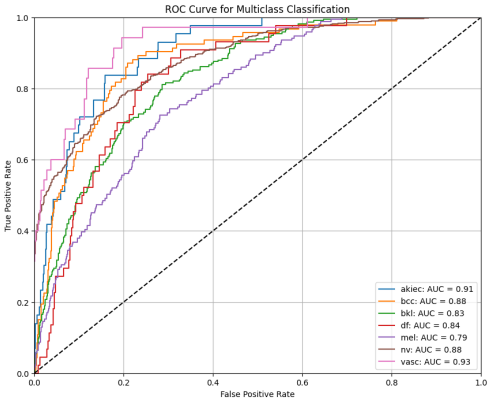


Fig. 2. ROC curve for Bayesian Optimization.

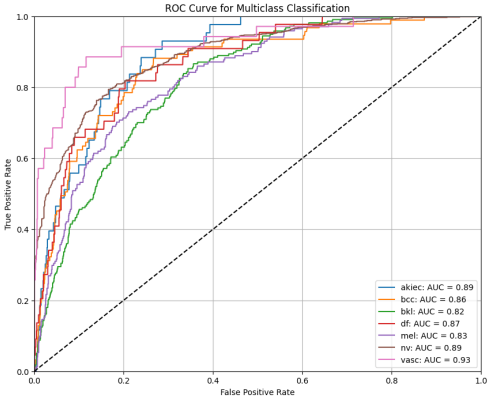


Fig. 3. ROC curve for Random Search.

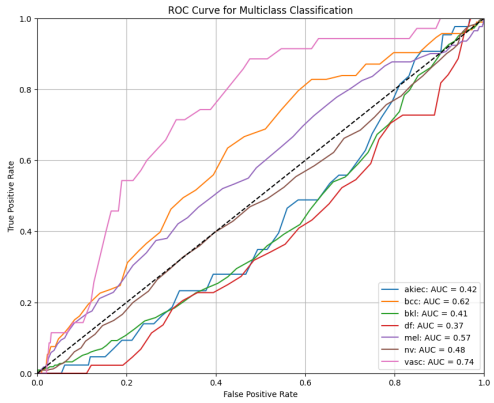


Fig. 4. ROC curve for SMAC3.

IV. Future Work

There are several directions in which future research could further improve the performance and applicability of this work in real-world medical settings. Exploring and fine-tuning pre-trained models could be an important next step. Models such as VGG and Vision Transformer [16], among others, may offer better performance for skin lesion classification after being fine-tuned on a dataset like HAM10000.

Deploying the model in a real-time environment would also be valuable. Given the medical context, it would be crucial to assess how well the model performs under real-world conditions, where images may vary in quality, lighting, or angle. Understanding how these conditions affect the model's performance is an important step toward its practical deployment.

A comparison with specialized networks, such as FixCaps, could help identify the limitations of more general models. This comparison could reveal areas where smaller networks like ours may fall short and guide future improvements in performance for skin lesion classification.

Few-shot learning techniques could be explored to address the challenge of acquiring large labeled datasets in medical domains. Investigating how the model can perform well with limited labeled samples would help overcome the data scarcity issue in medical imaging.

While this study focused on a relatively simple convolutional neural network, exploring deeper architectures could also be a promising avenue. Increasing the depth of the network while maintaining simplicity in the architecture may improve the model's ability to capture more intricate features in skin lesion images, striking a balance between model complexity and performance without significantly increasing computational cost.

This study also focused on Bayesian Optimization, Random Search, and SMAC3 optimization techniques due to hardware limitations. Future research could investigate other, more complex hyperparameter optimization methods, such as swarm or evolutionary algorithms, to explore whether they could further improve the model's performance.

V. Conclusion

This study explored the impact of various hyperparameter optimization techniques on the performance of a simple Convolutional Neural Network (CNN) for skin lesion classification. While Bayesian Optimization and Random Search methods achieved modest improvements over the baseline model, the performance gains were limited. Notably, Bayesian Optimization improved the F1 score, demonstrating better precision-recall balance, while SMAC3 underperformed. These results highlight the importance of hyperparameter tuning but also suggest that more significant performance gains may require architectural adjustments.

References

[1] American Cancer Society, *Cancer Facts & Figures 2025*. Atlanta: American Cancer Society, 2025.

[2] M. J. Ali, M. Essaid, L. Moalic, and L. Idoumghar, "A review of automl optimization techniques for medical image applications," *Computerized Medical Imaging and Graphics*, vol. 118, p. 102441, 2024.

[3] Z. Lan, S. Cai, X. He, and X. Wen, "Fixcaps: An improved capsules network for diagnosis of skin cancer," *IEEE Access*, vol. 10, pp. 76261–76267, 2022.

[4] H. Basak, R. Kundu, and R. Sarkar, "Mfsnet: A multi focus segmentation network for skin lesion segmentation," *Pattern Recognition*, vol. 128, p. 108673, Aug. 2022.

[5] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft-attention improves skin cancer classification performance," 2021.

[6] A. Romero Lopez, X. Giro-i Nieto, J. Burdick, and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," in *2017 13th IASTED International Conference on Biomedical Engineering (BioMed)*, pp. 49–54, 2017.

[7] T. Nguyen, V. H. Ho, and P. Do, "Optimizing deep learning for skin disease classification: Leveraging bayesian hyperparameter tuning and top-k accuracy metrics," in *Intelligent Systems and Data Science* (N. Thai-Nghe, T.-N. Do, and S. Benferhat, eds.), (Singapore), pp. 98–113, Springer Nature Singapore, 2025.

[8] F. Golnoori, F. Z. Boroujeni, and A. Monadjemi, "Metaheuristic algorithm based hyper-parameters optimization for skin lesion classification," *Multimedia Tools and Applications*, vol. 82, pp. 25677–25709, July 2023.

[9] P. Tschandl, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," 2018.

[10] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," 2012.

[11] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, *et al.*, "Kerastuner." https://github.com/keras-team/keras-tuner, 2019.

[12] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, p. 281–305, Feb. 2012.

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[14] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Learning and Intelligent Optimization* (C. A. C. Coello, ed.), (Berlin, Heidelberg), pp. 507–523, Springer Berlin Heidelberg, 2011.

[15] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhkopf, R. Sass, and F. Hutter, "Smac3: A versatile bayesian optimization package for hyperparameter optimization," *Journal of Machine Learning Research*, vol. 23, no. 54, pp. 1–9, 2022.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

Autism Spectrum Disorder Classification Using Functional Brain Connectivity

1st Jonatan Mileski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
jonatan.mileski@students.finki.ukim.mk

2nd Ilinka Ivanoska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
ilinka.ivanoska@finki.ukim.mk

***Abstract*—Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects the behavior, social communication, and social interaction of individuals. To date, there are no known biomarkers for diagnosing ASD, and the current diagnosis is based primarily on reported and observed behavior of tested individuals. This paper explores the potential of machine learning (ML) and deep learning (DL) in classifying ASD and typically developing (TD) individuals using resting-state fMRI (rs-fMRI). Specifically, functional connectivity patterns were derived from the Automated Anatomical Labeling (AAL) atlas, and multiple classification approaches were evaluated, including Deep Neural Networks (DNN), Graph Convolutional Networks (GCN), GraphSAGE, and Principal Component Analysis (PCA) combined with DNN. Among these, DNN demonstrated the highest classification performance, achieving an accuracy of 71%. These results suggest that even a reduced set of brain regions can provide meaningful information for ASD detection, yielding competitive results compared to studies using atlases with a larger number of regions. This study highlights the promise of deep learning in neuroimaging-based ASD classification and underscores the need for further research to improve accuracy and interpretability.**

***Index Terms*—Autism Spectrum Disorder, rs-fMRI, AAL, ABIDE-II, Deep Learning, Machine Learning, Functional Connectivity**

I. Introduction

Autism Spectrum Disorder (ASD) appears as a consequence of neural reorganization and early brain development alternations. It is characterized by repetitive sensory-motor behaviors and lack of social interaction and communication, although affected individuals can vary in numerous ways, often caused by additional comorbidities, resulting in a spectrum of symptoms [1]. Currently, ASD is diagnosed based on reported and observed behavior analysis, since there are no known generalized biomarkers indicating the disorder, although recent studies have suggested potential biomarkers, including genetic analysis, structural and functional connectivity patterns [2]. Recently, machine learning (ML) approaches and deep learning (DL) approaches have been reported as rather successful in highly accurate prediction of ASD based on the functional and structural patterns. Searching potential biomarkers for ASD detection has been an active topic in neuroimaging for a longer time. Several studies have shown success in the classification of ASD and typically developing (TD) samples, therefore suggesting potential biomarkers. Most studies use the Autism Brain Imaging Data Exchange (ABIDE) dataset [3]. For instance, J. Krajevski et al. in [4] explore the potential of traditional ML approaches and convolutional neural networks (CNNs) for ASD classification using the ABIDE-II dataset. For the ML methods, they suggest data transformation based on histogram calculations for dimensionality reduction, whereas for the CNNs they employ 3D-to-2D and 4D to-3D data transformation. Reported results show that Extra trees (ExT) achieved highest accuracy of 71.4%, both in structual and functional data.

In [5], five most widely employed ML and DL approaches are used for experiments with the structural volumetric measures, functional connectivity matrices, and phenotypic information from the ABIDE-II dataset. Experiment were conducted with fully connected networks (FCN), auto-encoder plus a fully connected network (AE-FCN), graph convolution networks (GCN), edge-variational graph convolution networks (EV-GCN) and support vector machine (SVM) models. Reported results of all methods showed similar classification metrics, while ensemble learning achieved the highest accuracy of 72.2%, and AUC score of 0.78 with GCN classifiers.

A. S. Heinsfeld et al. in [6] showed that using the rs-fMRI data from the ABIDE-II dataset with ROI functional connectivity extracted using the CC200 atlas, a deep neural network (DNN) learnt to classify ASD samples with an accuracy of 70% with 10-fold cross validation.

Similarly, X. Yang et al. in [7] experimented with DNN architectures using the rs-fMRI data from the ABIDE-II dataset. ROIs were extracted using the CC400 atlas, based on its leading score in comparison made in their previous study with other atlases. Their most successful DNN achieved an accuracy of 75.27%. The hidden layers consisted of 128 and 64 units, in the first and the second hidden layer, respectively.

Perhaps, most similar to our work is presented in [8]. The presented experiments also base on functional connectivity study using ABIDE-II dataset. Experiments based on different atlases, including AAL are conducted. Models used in this study are support vector machines, logistic regression and ridge. For finding the optimal parameteres, authors use cross-validation grid search. The optimal parameters achieved an accuracy of 71.98%.

This study explores how different ML and DL architectures

trained with different data transformations will affect the classification metrics of ASD and TD classification. We use the second iteration of the publicly-available ABIDE dataset. All our experiments are based on the regions of interest (ROI) functional connectivity correlations, obtained from the Automated anatomical labeling (AAL) [9] atlas, which contains 116 brain regions. We present and discuss the obtained classification metrics results.

The paper is organized as follows. In section II are described the dataset used, data preprocessing steps, data transformation techniques and the experimental methodology used. Afterwards, section II describes the performed experiments, obtained results and discussion. At the end, section IV provides the concluding remarks of our study.

II. Materials and Methods

Figure 1 illustrates the overall system architecture used in this study, detailing the data flow from preprocessing to classification.

A. Dataset

The process begins with the acquisition of functional magnetic resonance imaging (fMRI) data from the ABIDE-II dataset, which comprises 19 sites and two longitudinal collections, including both ASD and TD subjects. The total number of subjects in the dataset is 1,114 (521 ASD and 593 TD). For the purposes of this study, a subset of the ABIDE-II dataset with *N*=991 (ASD=474, 48% of the dataset subset; TD=517, 52% of the dataset subset) was extracted from 16 sites and the 2 longitudinal collections. From subjects with more than one resting-state session, only the first session was taken into consideration, and if missing, the subject was discarded. Table I reports the number of subjects used from each site.

TABLE I
Subjects count per site

Site	# of Subjects
BNI_1	58
EMC_1	54
ETH_1	37
GU_1	75
IP_1	55
IU_1	40
KKI_1	211
KUL_3	28
NYU_1	78
NYU_2	27
OHSU_1	93
SDSU_1	58
TCD_1	42
UCD_1	32
UCLA_1	32
USM_1	33
UPSM_Long	17
UCLA_Long	21

B. Data preprocessing

The preprocessing of the subjects was done with MATLAB R2024b and CONN toolbox 22.v2407 [10]. The following preprocessing steps were performed: subject motion estimation and correction; slice-timing correction; outlier detection; direct segmentation and MNI-space normalization; smoothing with 8mm full width at half maximum (FWMH) kernel and [0.008, 0.09] band-pass filtering. In all our experimts we use the AAL atlas ROI to ROI functional connectivity matrices with Pearson coefficient as correlation measurement. Since AAL contians 116 brain regions of interest, the matrices are 116x116. The preprocessing pipeline was run on Azure DS4_v3 virtual machine with 4 virtual CPUs and 16GB RAM.

C. Data transformation

To evaluate the impact of different preprocessing techniques on classification performance, various data transformations were applied to the connectivity matrices. Two normalization techniques were considered: min-max normalization and z-score normalization, both applied at the individual subject level. Additionally, experiments were conducted to investigate the effect of thresholding on correlation values. Specifically, certain experiments replaced correlation values exceeding a predefined threshold with either 1 or -1, based on the sign of the correlation, thereby converting the connectivity matrices into binary representations. These transformations aimed to explore how different preprocessing strategies influence model performance in ASD classification.

D. Methodology

This study employs the following classical and state-of-the-art ML and DL methods in the conducted experiments:

- **DNN** [11] - The architecture that we use employs an input layer of 6,670 (the total number of ROI-ROI correlations). Then the input is forwarded to a fully connected 40-units layer with ReLU activation. Then a hidden layer of 60 units with dropout 0.5 and ReLU activation follows, after which a fully connected layer of 20 hidden units with ReLU activation follows. Finally, for class prediction there is 1 unit with Sigmoid activation.
- **GCN** [12] - The input to the model is a graph representation of each subject, individually, where nodes correspond to ROIs and edges represent functional connectivity between them. We use two GraphConv layers with sigmoid activation, and global pooling afterwards. Finally, a fully connected layer is used for class prediction.
- **GraphSAGE** [13] - Same as in GCN, the input to this model are the subject-specific graphs representing the ROIs functional connectivity. Two SAGEConv layers with dropout 0.3 and hyperbolic tangent as activation function are used. Furthermore, global max pooling is applied after which fully connected layer for class prediction follows.
- **Principle Component Analysis (PCA) + DNN** - This approach used PCA for dimensionality reduction with 50 components, which drastically lowers the dimensionality of the original 6,670 features. Afterwards, the newly obtained features are passed to a DNN with the same architecture as previously specified in this subsection.

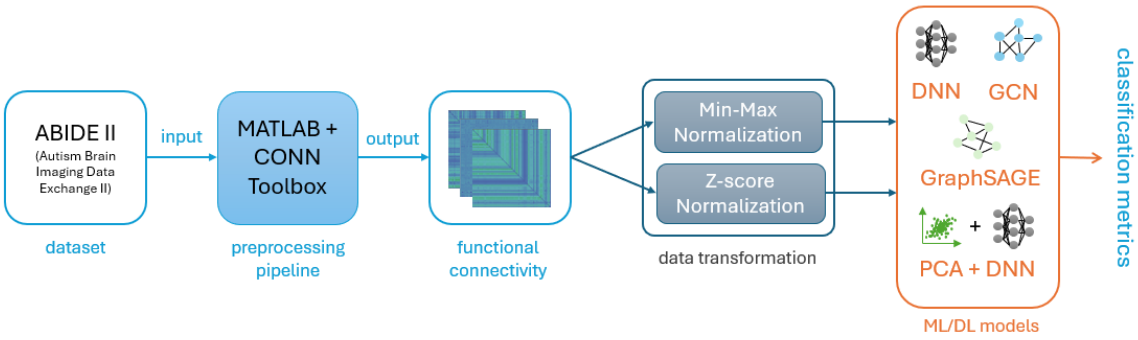


Fig. 1. Project architecture

All models were trained on an NVIDIA Tesla T4 GPU to leverage hardware acceleration for deep learning computations. The architectural workflow depicted in Fig. 1 provides a comprehensive overview of the data processing pipeline, from raw fMRI acquisition to model inference, ensuring an efficient and systematic approach to ASD classification.

III. Experiments, Results and Discussion

Using the different data transformations mentioned in the previous section, we performed DNN, GCN, GraphSAGE and Principle Component Analysis (PCA)-based DNN experiments. Additionally, 4-fold stratified cross-validation (CV) [14] technique was used for all the mentioned approaches. Based on the dataset size, and on several experiments with different values for the number of folds, we concluded that 4-fold CV is the most appropriate in this case. For the DNN and the PCA plus a DNN approaches, we mapped the half above the diagonal of the matrices into vectors of size 6,670. For GCN and GraphSAGE, PyTorch [15] graphs were created with nodes representing the ROIs and edges representing the high correlation values. In all the experiments, DNN showed highest performance, with results achieving 71% accuracy in the best case. The most successful was the case when z-score normalization was used and a threshold of 0.8, meaning that all the ROI-ROI correlations whose absolute value was smaller than 0.8 were substituted with 0, while all the other correlations larger or equal to the threshold were mapped to the absolute value of their current correlation value after the z-score normalization. Table II shows the results of all different methods used with different data transformations. Fig. 2 shows the ROC curve of the best DNN case.

Overall, the other 3 classification approaches performed worse than DNN, when taking into consideration the four metrics: accuracy, precision, recall and f1-score.

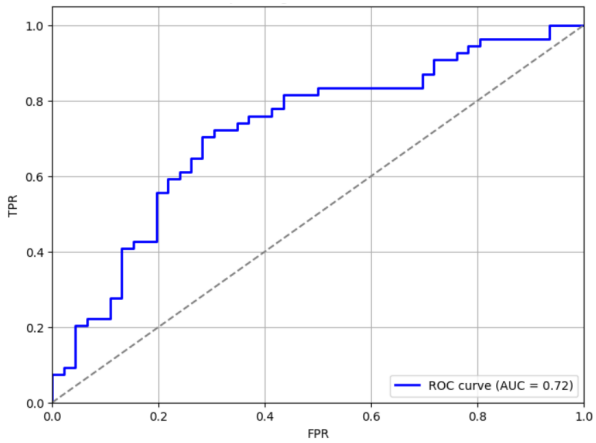


Fig. 2. ROC curve of DNN with 4-fold CV

IV. Conclusion

The transformation of ROI-to-ROI functional connectivity matrices into sparse representations using thresholding and absolute correlation values significantly improved ASD vs. typical controls classification performance. From the observed results, z-score normalization emerged as the most effective data transformation technique, while the DNN architecture outperformed all other models.

Furthermore, the experiments confirmed the necessity of K-fold CV, particularly given the limited dataset size, where overfitting could otherwise lead to degraded test performance. The inferior results of GCN and GraphSAGE suggest that these models are less suited for this problem due to their reliance on local neighborhood aggregation, which does not account for the positional dependencies inherent in ROI-ROI connectivity matrices. Regarding the PCA+DNN approach results, we believe that with only 50 components instead of 6,670 we lose significant information, even though there was a case when the recall was 79,6%, but the other metrics were lower than the best results obtained with DNN. As a result, DNN-based approaches remain the most promising solution for ASD classification using functional brain connectivity data.

Future research directions could focus on improving the classification performance by exploring advanced deep learning architectures and alternative data representations. Additionally, incorporating multimodal neuroimaging data, such as structural MRI or diffusion tensor imaging (DTI), alongside

TABLE II
PERFORMANCE OF DIFFERENT DATA TRANSFORMATIONS AND APPROACHES

Transformation	Metric	Methods			
		DNN	PCA+DNN	GCN	GraphSAGE
Min-max normalization	Accuracy	0.580	**0.600**	–	–
	Recall	**0.703**	0.648	–	–
	Precision	0.593	**0.625**	–	–
	F1-score	**0.644**	0.636	–	–
Min-max normalization in range [-1, 1]	Accuracy	**0.620**	0.540	–	–
	Recall	0.703	**0.759**	–	–
	Precision	**0.633**	0.554	–	–
	F1-score	**0.667**	0.641	–	–
Min-max normalization in range [-1, 1] and mapping all values whose absolute value is less than 0.8 to 0	Accuracy	0.570	**0.650**	0.504	0.528
	Recall	0.611	**0.778**	0.504	0.528
	Precision	0.600	**0.646**	0.505	0.534
	F1-score	0.605	**0.706**	0.503	0.521
Min-max normalization in range [-1, 1] and mapping all values whose absolute value is less than 0.8 to 0, while the rest of the values are mapped to 1	Accuracy	0.600	**0.640**	0.508	0.496
	Recall	0.629	**0.796**	0.508	0.496
	Precision	0.629	**0.632**	0.509	0.499
	F1-score	0.629	**0.705**	0.508	0.490
Min-max normalization in range [-1, 1] and mapping all values whose absolute value is less than 0.8 to 0, while the others are mapped to 1 or -1 depending on their original sign	Accuracy	**0.600**	0.560	0.496	0.528
	Recall	**0.703**	**0.703**	0.496	0.528
	Precision	**0.612**	0.576	0.495	0.529
	F1-score	**0.655**	0.633	0.495	0.528
Z-score normalization and mapping all values whose absolute value is less than 1 to 0	Accuracy	0.580	0.560	0.516	**0.568**
	Recall	0.648	**0.722**	0.516	0.569
	Precision	**0.603**	0.574	0.520	0.576
	F1-score	0.625	**0.639**	0.510	0.563
Z-score normalization and mapping all values whose absolute value is less than 0.8 to 0, while the others are mapped to 1 or -1 depending on their original sign	Accuracy	**0.690**	0.540	0.532	0.548
	Recall	**0.648**	0.556	0.532	0.548
	Precision	**0.744**	0.577	0.541	0.551
	F1-score	**0.693**	0.566	0.519	0.547
Z-score normalization and mapping all values whose absolute value is less than 0.8 to 0, while the other values are mapped to their absolute value	Accuracy	**0.710**	0.630	0.512	0.524
	Recall	**0.703**	**0.703**	0.512	0.524
	Precision	**0.745**	0.644	0.515	0.525
	F1-score	**0.723**	0.672	0.509	0.524

The best result of each row is in bold

fMRI-derived functional connectivity could provide complementary features that enhance classification accuracy. Fusion techniques combining multiple modalities may offer a more comprehensive characterization of ASD-related brain patterns. Further improvements could also be achieved through more sophisticated feature engineering techniques, such as graph embedding methods that better preserve spatial relationships between ROIs. Expanding the dataset by integrating additional publicly available neuroimaging datasets or applying data augmentation techniques specific to fMRI could help mitigate overfitting and improve generalizability.

REFERENCES

[1] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *Lancet*, vol. 392, pp. 508–20, 2018.

[2] L. Zwaigenbaum and M. Penner, "Autism spectrum disorder: advances in diagnosis and evaluation," *Bmj*, vol. 361, 2018.

[3] A. Di Martino, D. O'connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, J. H. Balsters, L. Baxter, A. Beggiato, S. Bernaerts, *et al.*, "Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii," *Scientific data*, vol. 4, no. 1, pp. 1–15, 2017.

[4] J. Krajevski, I. Ivanoska, K. Trivodaliev, S. Kalajdziski, and S. Gievska, "An exploration of autism spectrum disorder classification from structural and functional mri images," in *International Conference on ICT Innovations*, pp. 175–189, Springer, 2022.

[5] Y. Dong, D. Batalle, and M. Deprez, "A framework for comparison and interpretation of machine learning classifiers to predict autism on the abide dataset," *Human Brain Mapping*, vol. 46, no. 5, p. e70190, 2025.

[6] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the abide dataset," *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018.

[7] X. Yang, P. T. Schrader, and N. Zhang, "A deep neural network study of the abide repository on autism spectrum classification," *International journal of advanced computer science and applications*, vol. 11, no. 4, 2020.

[8] X. Yang, M. S. Islam, and A. A. Khaled, "Functional connectivity magnetic resonance imaging classification of autism spectrum disorder using the multisite abide dataset," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1–4, IEEE, 2019.

[9] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[10] S. Whitfield-Gabrieli and A. Nieto-Castanon, "Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks," *Brain connectivity*, vol. 2, no. 3, pp. 125–141, 2012.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[13] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[14] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.

[15] A. P. et al., "Pytorch: An imperative style, high-performance deep learning library," 2019.

Performance Analysis of Open-Source Large Language Models: Accuracy, Efficiency, and Response Variability

1th Vesna Pop-Dimitrijoska Koteska
LOKA, Los Altos, CA, 94022, USA
vesna@loka.com

2nd Ana Markovska
LOKA, Los Altos, CA, 94022, USA
ana.markovska@loka.com

3rd Andrej Petkovikj
LOKA, Los Altos, CA, 94022, USA
andrej.petkovikj@loka.com

4th Marija Maneva
LOKA, Los Altos, CA, 94022, USA
marija.maneva@loka.com

5st Ema Pandilova
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
ema.pandilova@finki.ukim.mk

6th Pance Ribarski
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
pance.ribarski@finki.ukim.mk

7th Bojan Ilijoski
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
bojan.ilijoski@finki.ukim.mk

***Abstract*—This paper evaluates the performance of various Large Language Models (LLMs) across diverse tasks, focusing on their efficiency, accuracy, and computational requirements. Models such as OpenChat-3.5, TinyLlama-1.1B, Phi-2, and Orca Mini V3 were assessed using a structured set of questions spanning domains like geography, mathematics, and general knowledge. The study analyzed the impact of different temperature settings on response variability and reasoning capabilities. Findings indicate that lower inference times do not necessarily correlate with higher accuracy; smaller models like TinyLlama-1.1B were efficient but struggled with complex reasoning tasks. Higher temperature settings (0.9) led to more diverse responses but also increased hallucination rates in complex queries. These insights underscore the importance of selecting appropriate models based on specific task requirements and highlight the potential of fine-tuning smaller models for domain-specific applications to balance efficiency and accuracy.**

***Index Terms*—Large Language Models, Natural Language Processing, Inference Time, Model Performance, AI Evaluation**

I. Introduction

Large Language Models (LLMs) have become a central focus in artificial intelligence research, demonstrating significant advancements in natural language processing. As these models continue to evolve, evaluating their behavior, strengths, and limitations is essential for understanding their real-world applicability [1].

This study explores the performance of multiple LLMs, including Sonya-7B, OpenChat-3.5, TinyLlama-1.1B, Phi-2, and Orca Mini V3, by deploying them locally with llama_cpp. The evaluation involved running the models under different temperature settings and analyzing their responses to a structured set of questions. These questions spanned multiple domains, including geography, food, movies, and mathematics, and were categorized into easy, medium, and hard difficulty levels.

To ensure an objective assessment, a 5-point scoring system was implemented to measure the accuracy and relevance of model responses. Additionally, computational factors such as processor performance were taken into account to maintain fairness in evaluating response time. By systematically analyzing the models' outputs, this research provides valuable insights into their capabilities and limitations, highlighting patterns in model performance across different contexts.

The paper is organized as follows: In the second section we provide an overview of Large Language Models (LLMs), describing their architecture and operational principles. The third section details the specific models used in this study, outlining their characteristics and computational requirements. In the fourth section, we present the experimental setup, including the methodology used to evaluate model performance across different tasks and temperature settings. The fifth section discusses the results, analyzing model accuracy, inference time, and response variability across difficulty levels. The concluding remarks are provided in the final section.

II. Large Language Models (LLMs)

Large language models largely represent a class of deep learning architectures called transformer networks. A transformer model is a neural network that learns context and meaning by tracking relationships in sequential data, like the words in this sentence. A transformer is made up of multiple transformer blocks, also known as layers. For example, a transformer has self-attention layers, feed-forward layers, and normalization layers, all working together to decipher input to predict streams of output at inference. The layers can be stacked to make deeper transformers and powerful language models [2]. Approaches for different neural architectures include decoder-only, encoder-only, and encoder-decoder architectures.

A. Encoder-Only

An encoder is a function that reads a sequence of tokens $x = x_0, \ldots, x_m$ and produces a sequence of vectors $H = h_0, \ldots, h_m$. A typical approach to encoder pre-training is to combine the encoder with some output layers to receive supervision signals that are easier to obtain.

A common architecture for pre-training an encoder-based model is to add a Softmax layer on top of the encoder. Here, $p_i^{W,\theta}$ is the output distribution at position i. We use Softmax^W to denote that the Softmax layer is parameterized by W, that is,

$$\text{Softmax}^W(H) = \text{Softmax}(H \cdot W)$$

For notational simplicity, we will sometimes drop the superscripts W and θ affixed to each probability distribution.

These models are typically suitable for tasks that require sentiment analysis, classification, or general language understanding [2].

B. Encoder-Decoder

In NLP, Encoder-Decoder architectures are often used to model sequence-to-sequence problems, such as machine translation and question answering. Except for these sequence-to-sequence we can also consider text as both the input and output of a problem, and so we can directly apply Encoder-Decoder models. For example, given a text, we can ask a model to output a text describing the sentiment of the input text, such as positive, negative, and neutral. We first train an Encoder-Decoder model to gain general-purpose knowledge of language via self-supervision. This model is then fine-tuned for specific downstream tasks using targeted text-to-text data [2]. The general architecture of such models is illustrated in Fig. 1.

C. Decoder-Only

The Transformer-decoder architecture for language modeling. The central components are L stacked Transformer blocks, each comprising a self-attention sub-layer and an FFN sub-layer. To prevent the model from accessing the right-context, a masking variable is incorporated into self-attention. The

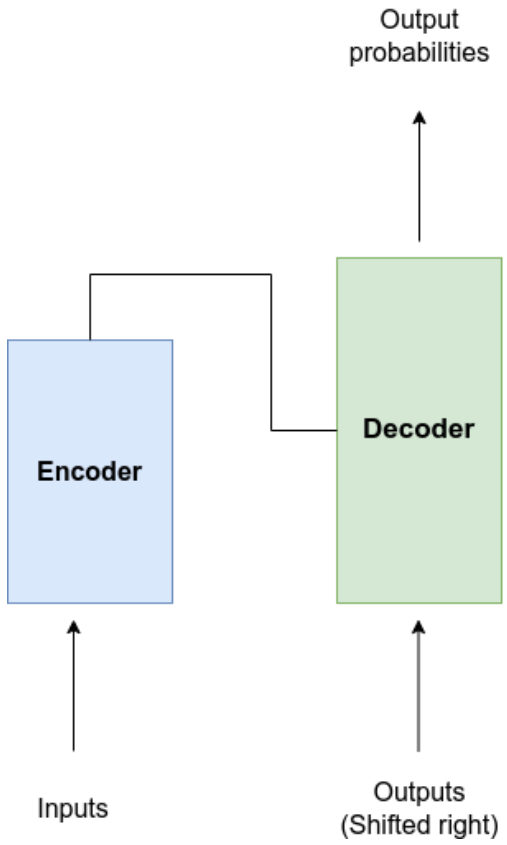


Fig. 1. Illustration of the Encoder-Decoder Architecture

output layer uses a Softmax function to generate a probability distribution for the next token, given the sequence of previous tokens. During inference, the model takes the previously predicted token to predict the next one, repeating this process until the end of the sequence is reached [2].

D. Creating Powerhouse LLMs

Large Language Models can be trained using supervised, unsupervised, and reinforcement learning. Supervised learning as a way of training the model learns from the labeled data and tries to make a prediction on newly seen data. With unsupervised learning, models can find previously unknown patterns in data using unlabelled datasets. This also eliminates the need for extensive data labeling, which is one of the biggest challenges in building AI models. Unlike supervised learning, which depends on labeled data, and unsupervised learning, which uncovers patterns in unlabeled data, reinforcement learning emphasizes learning through direct feedback via trial and error [3]. In reinforcement learning (RL), there are six key components:agent, environment, state, action, reward, and policy. To apply RL for fine-tuning large language models (LLMs), the first step is to map these components to the LLM framework. LLMs are highly proficient at next-token prediction, where they take a sequence of tokens as input and predict the next token based on the given context. From an

RL perspective, we can view the LLM itself as the policy. The current textual sequence represents the state, and based on this state, the LLM generates an action—the next token. This action updates the state, creating a new state that incorporates the newly added token. After generating a complete textual sequence, a reward is determined by assessing the quality of the LLM's output using a pre-trained reward model [3]. For truly creating/crafting a powerhouse LLM the key is to use the best of both worlds or in our case combining multiple training methods together.

III. Overview of Large Language Models Used

A. Sonya-7B

Sonya-7B is a high-performing 7B parameter model built as a merge of several specialized open-source models, designed for efficient local inference and high accuracy in general-purpose tasks. In our experiments, Sonya-7B consistently achieved top scores across a range of question difficulties, including factual and reasoning-based prompts. The model's performance suggests that it has been optimized for balanced generation, with a strong ability to produce coherent and contextually relevant responses even under resource constraints. Sonya-7B was deployed using the llama_cpp backend in GGUF format and demonstrated favorable inference times compared to other models in the 7B parameter range. Although the precise training data and fine-tuning methodology are not publicly disclosed, the model's outputs indicate strong generalization capabilities across diverse domains.

B. Openchat-3.5

We also utilized the OpenChat 3.5 model, locally hosted in GGUF format, to evaluate its conversational response generation capabilities within our experimental framework. It is an open-source conversational language model designed to enhance dialogue generation through advanced fine-tuning techniques. Unlike traditional supervised fine-tuning (SFT) methods, OpenChat-3.5 incorporates a novel Conditioned Reinforcement Learning Fine-Tuning (C-RLFT) approach, which leverages mixed-quality training data to improve response generation. By integrating coarse-grained reward signals, the model balances efficiency and response quality without the need for expensive human preference annotations. [4].

C. Tinyllama-1.1b

TinyLlama-1.1B is a compact, open-source language model designed to optimize performance within constrained computational environments. Its architecture follows the LLaMA model family, incorporating optimizations such as grouped-query attention, SwiGLU activation, and rotary positional embeddings to enhance efficiency [5]. In our research, we employed TinyLlama-1.1B to evaluate its capabilities in generating domain-specific responses while maintaining a lightweight computational footprint.

D. Phi-2

Phi-2, originally developed by Microsoft, is a streamlined small language model requiring fewer computational resources compared to larger models like GPT-3.5. In our research, we utilized Phi-2 to explore its capabilities in generating domain-specific responses and evaluated its performance under various experimental conditions. The model's lightweight architecture and efficient fine-tuning methods made it particularly suitable for local deployment and testing within our framework [6].

E. Orca_mini_v3_7b

The orca_mini_v3_7b is a compact language model developed by Pankaj Mathur in 2023 [7]. Designed to perform efficiently on systems with limited hardware resources, it offers strong performance on language model leaderboards despite its smaller size.

F. Llama-2-chat-7B

Llama-2-Chat-7B is a version of the Llama 2 model that has been fine-tuned to handle conversations more effectively. It uses a training method called Reinforcement Learning with Human Feedback (RLHF) to improve how well it follows instructions and responds to questions. This model has been designed to be safer and more useful, making it a strong alternative to popular AI chatbots [8]. In our research, we tested Llama-2-Chat-7B to see how well it generates clear, relevant, and accurate responses across different topics.

G. Zephyr-7b-beta

Zephyr-7B-Beta is a fine-tuned version of the Mistral-7B model, optimized for intent alignment using distilled direct preference optimization (dDPO). This approach eliminates the need for costly human feedback by leveraging AI-generated preference rankings, significantly improving the model's ability to follow user instructions accurately. Zephyr-7B-Beta has achieved state-of-the-art performance among 7B-parameter models, even surpassing larger models like LLaMA-2-Chat-70B on chat-based benchmarks [9].

H. Mistral-7b-openorca

Mistral-7B-OpenOrca is an enhanced version of the Mistral-7B model, fine-tuned with high-quality instruction-following data to improve its reasoning, comprehension, and alignment with human intent. By leveraging techniques such as explanation tuning and large-scale instruction tuning, OpenOrca enables Mistral-7B to perform more effectively on complex reasoning tasks, including multi-step problem-solving and knowledge-intensive queries. This model incorporates a diverse dataset, including step-by-step explanations generated by larger models like GPT-4, allowing it to better emulate the reasoning process of more advanced language models. As a result, Mistral-7B-OpenOrca achieves competitive performance on challenging benchmarks such as BigBench-Hard and AGIEval, reducing the gap between smaller open-source models and proprietary systems while maintaining efficiency for deployment in resource-constrained environments [10].

I. Mistral-7B

Mistral-7B is a causal language model designed with a transformer decoder-only architecture, prioritizing both efficiency and strong performance in natural language processing tasks. It employs Grouped-Query Attention (GQA), which optimizes key-value head sharing to maintain a balance between computational cost and model capacity, enabling it to process queries more efficiently while preserving the quality of responses. Additionally, Sliding Window Attention (SWA) enhances the model's ability to capture local context effectively, reducing computational complexity during inference while still ensuring that relevant dependencies are maintained across tokens [11].

J. Starling-LM-7B-alpha

Starling-LM-7B-alpha is a high-performance 7B chat model developed through fine-tuning Openchat-3.5 using the Nectar dataset, a high-quality preference dataset generated through GPT-4-ranked responses. The training approach leverages an internal pairwise rating technique, allowing the model to efficiently compare response quality without requiring individual pairwise evaluations, thereby optimizing ranking decisions at scale. As a result, Starling-LM-7B demonstrates state-of-the-art performance on MT-Bench, AlpacaEval, and human evaluation metrics, establishing itself as a leading model in chat-based AI interactions [12].

IV. Experimental Setup

For the evaluation of the LLMs, a systematic framework was established to assess the performance across multiple dimensions. This approach builds upon established evaluation frameworks in LLM research that emphasize systematic assessment across multiple dimensions [13]. A diverse set of open-source LLMs were selected, including Llama-2, Mistral, Phi-2, TinyLlama, Zephyr, Starling-LM, and Orca Mini for comparison. All models were implemented using the llama_cpp Python library with consistent initialization parameters where possible, including context length (n_ctx) values ranging from 2048 to 32768, and thread counts optimized for the local hardware configuration.

To evaluate model performance under different inference conditions, each model was tested with two distinct temperature settings: a deterministic setting (temperature = 0.0) and a more creative setting (temperature = 0.9). Similar to recent work in systematic LLM evaluation, our approach employs diverse test cases to comprehensively assess model capabilities across different tasks and knowledge domains. The standardized testing protocol that was developed uses a predefined set of prompts covering various knowledge domains and task types, including factual questions, recipe generation, and analytical reasoning [14].

Following best practices in LLM evaluation literature, we incorporated both qualitative accuracy assessments and quantitative performance measurements to provide a holistic view of model capabilities [14]. Each model processed identical prompts with timing data collected using Python's time.time() function to measure end-to-end response generation. All experiments were conducted on each local machine to ensure fair comparison across models, with no GPU acceleration (n_gpu_layers=0) to establish a CPU-only baseline. Due to the variation in model sizes and their computational requirements, we tested smaller LLMs on less powerful hardware and larger LLMs on more capable configurations. This reflects realistic deployment scenarios where lightweight models are preferred for edge or resource-constrained environments, while larger models are typically hosted on higher-end machines. While uniform testing on the same hardware could provide a fully controlled comparison, our approach offers a more practical evaluation of each model's performance relative to its typical usage context. The results were systematically recorded in a comparison spreadsheet for subsequent analysis and classification of the model.

Later on, LangChain and LlamaIndex were employed to facilitate the setup and querying of large language models (LLMs) from a local machine. LangChain is a comprehensive framework designed to streamline the development of applications utilizing large language models, offering tools and abstractions that simplify interactions with LLMs, manage prompt engineering, and integrate external data sources effectively [15] . LlamaIndex complements this by providing an efficient mechanism to extend the context window of language models, enabling structured indexing and retrieval of external knowledge bases to enhance model responses and accuracy [16]. Together, these tools significantly accelerate the deployment and querying process for sophisticated LLM-based applications.

V. Results and Discussion

The evaluation of multiple open-source LLMs revealed distinct variations in their performance across different difficulty levels, domains, and computational efficiency. One of the key observations is that model responses were highly dependent on the temperature setting, impacting both response accuracy and creativity. The models performed consistently well on factual, well-documented topics such as geography but demonstrated significant variability in reasoning-based or niche knowledge areas like movie-related queries.

A. Performance Across Models

In the study presented in Tables I and II, we compare the performance of various large language models (LLMs) under different conditions. These tables show each model's average score and how fast they process information, measured by inference time, along with the hardware configuration. When we look at Table I, we see a wide range of inference times at a temperature setting of 0.9. For instance, the model called Mistral-7B-Instruct took the longest time, 1475 seconds, to complete tasks, which suggests it might need more powerful hardware or optimization to work faster. On the other end, TinyLlama-1.1b was the quickest, taking only 2 seconds to process tasks, showing it's very efficient. Despite these efficiency differences, scoring patterns remained relatively stable,

TABLE I
COMPARISON OF LLM PERFORMANCE BASED ON SCORE, INFERENCE TIME, AND HARDWARE CONFIGURATION (TEMP = 0.9)

No	LLM	Avg. Score (1–5)	Avg. Score: EASY Questions (1–5)	Avg. Score: MEDIUM Questions (1–5)	Avg. Score: HARD Questions (1–5)	Avg. Inference Time (s)	Hardware Configuration
1	Sonya-7B	4.7	4.5	5.0	4.5	135	Intel Core i5, 1.6 GHz, 8 GB RAM
2	openchat-3.5	4.5	5.0	5.0	3.7	42	Intel Core i5-8265U, 1.60 GHz, 16 GB RAM, 8 CPUs
3	Tinyllama-1.1b	3.1	4.5	2.0	3.3	2	Intel Core i7-1165G7, 2.80 GHz, 16 GB RAM, 8 CPUs
4	Phi-2	4.4	4.5	5.0	4.0	12	Intel Core i7-1165G7, 2.80 GHz, 16 GB RAM, 8 CPUs
5	Orca_mini_v3_7b	4.3	4.5	4.0	3.7	16	Intel Core i7-1165G7, 2.80 GHz, 16 GB RAM, 8 CPUs
6	llama-2-chat-7B	4.8	4.5	5.0	4.8	34	AMD Ryzen 5 3500U, 2.10 GHz, Radeon Vega Mobile Gfx, 8 GB RAM, 8 CPUs
7	zephyr-7b-beta	3.8	5.0	3.7	3.3	20	AMD Ryzen 5 3500U, 2.10 GHz, Radeon Vega Mobile Gfx, 8 GB RAM, 8 CPUs
8	mistral-7b-openorca	4.8	5.0	5.0	4.3	15	AMD Ryzen 5 3500U, 2.10 GHz, Radeon Vega Mobile Gfx, 8 GB RAM, 8 CPUs
9	mistral-7b-instruct	4.7	5.0	5.0	4.0	1475	Intel Core i7-1260P, 2.10 GHz, 16 GB RAM, 16 CPUs
10	starling-lm-7b-alpha	4.7	4.5	5.0	4.5	87	Intel Core i7-1260P, 2.10 GHz, 16 GB RAM, 16 CPUs

TABLE II
COMPARISON OF LLM PERFORMANCE BASED ON SCORE, INFERENCE TIME, AND HARDWARE CONFIGURATION (TEMP = 0.0)

No	LLM	Avg. Score (1–5)	Avg. Score: EASY Questions (1–5)	Avg. Score: MEDIUM Questions (1–5)	Avg. Score: HARD Questions (1–5)	Avg. Inference Time (s)	Hardware Configuration
1	Sonya-7B	4.9	4.5	5.0	5.0	93	Intel Core i5, 1.6 GHz, 8 GB RAM
2	openchat-3.5	4.8	5.0	4.7	4.8	35	Intel Core i5-8265U, 1.60 GHz, 16 GB RAM, 8 CPUs
3	Tinyllama-1.1b	3.4	3.5	3.7	3.3	2	Intel Core i7-1165G7, 2.80 GHz, 16 GB RAM, 8 CPUs
4	Phi-2	4.6	4.5	5.0	4.3	6	Intel Core i7-1165G7, 2.80 GHz, 16 GB RAM, 8 CPUs
5	Orca_mini_v3_7b	4.4	4.5	4.0	4.3	17	Intel Core i7-1165G7, 2.80 GHz, 16 GB RAM, 8 CPUs
6	llama-2-chat-7B	4.4	4.5	3.3	3.9	18	AMD Ryzen 5 3500U, 2.10 GHz, Radeon Vega Mobile Gfx, 8 GB RAM, 8 CPUs
7	zephyr-7b-beta	4.3	4.5	5.0	3.8	31	AMD Ryzen 5 3500U, 2.10 GHz, Radeon Vega Mobile Gfx, 8 GB RAM, 8 CPUs
8	mistral-7b-openorca	4.7	4.5	5.0	4.5	30	AMD Ryzen 5 3500U, 2.10 GHz, Radeon Vega Mobile Gfx, 8 GB RAM, 8 CPUs
9	mistral-7b-instruct	4.8	5.0	5.0	4.5	104	Intel Core i7-1260P, 2.10 GHz, 16 GB RAM, 16 CPUs
10	starling-lm-7b-alpha	4.7	4.5	5.0	4.5	71	Intel Core i7-1260P, 2.10 GHz, 16 GB RAM, 16 CPUs

with most models achieving scores between 3 (neutral) and 5 (strongly agree).

Similarly, in Table II, we see how these models performed at a lower temperature setting of 0.0. Interestingly, some models like the Mistral-7B-Instruct still took longer than most, 104 seconds, but this was much quicker compared to its performance at the higher temperature. This shows that changing conditions can significantly affect how fast these models work. Even though the processing times varied, the quality of the models' outputs, indicated by their scores, remained relatively high. Most models scored again between 3 to 5, which means they performed from neutral to very well in understanding and processing tasks.

B. Impact of Question Difficulty on Model Performance

To better understand how LLMs handle varying levels of complexity, questions were categorized into Easy, Medium, and Hard difficulty levels based on their cognitive requirements (Table III). The easy questions were generally answered accurately across all models, as they involved simple factual retrieval. Medium-difficulty questions, such as basic mathematical reasoning or ingredient-based queries, saw some variability, particularly in models with smaller architectures. Hard questions posed the greatest challenge, with hallucinated answers appearing frequently in movie-related and subjective recommendation-based queries. The "Reasoning" column explains why the question was classified at its respective difficulty level, based on anticipated challenges for small open-source LLMs rather than observed outputs.

C. Observations on Model Performance

- **Low inference time does not always correlate with better accuracy** – Some smaller models like TinyLlama-1.1B were highly efficient, but received lower scores on medium and hard difficulty questions due to limited reasoning capabilities.
- **Temperature plays a crucial role in response variability** – At higher temperatures (0.9), models produced more diverse responses, but with increased hallucination rates in complex queries.
- **Hardware constraints influence model selection** – The larger 7B models required significantly more computational resources, making efficiency trade-offs essential for real-world deployment.

In conclusion, these findings underscore the importance of selecting the appropriate model based on task requirements. Future work should explore fine-tuning small models for specialized domains to enhance their reliability while maintaining computational efficiency.

VI. CONCLUSION

The evaluation of multiple LLMs revealed distinct variations in performance across different domains and difficulty levels. One of the most significant findings was the tendency of models to struggle with movie-related questions, frequently generating hallucinated responses. This pattern suggests potential gaps in training data or reasoning mechanisms when handling specific types of knowledge.

The use of a 5-point evaluation scale ensured consistency in scoring, allowing for a structured comparison of accuracy and response quality. Additionally, consideration of hardware differences provided a more balanced perspective on inference speed.

Overall, the results highlight the strengths and weaknesses of various LLMs, emphasizing the importance of fine-tuning

TABLE III
QUESTION CLASSIFICATION BY DIFFICULTY LEVEL FOR SMALL OPEN-SOURCE LLMS

No	Questions	Difficulty Level	Reasoning
1	What is the capital city of Germany?	Easy	Simple factual knowledge, well-documented in training data
2	Which countries are located in Scandinavia?	Easy	Common geography, straightforward lookup
3	How much protein is in one egg?	Medium	Requires basic nutrition knowledge, but small LLMs may struggle with precise values
4	What drinks go into a mojito?	Medium	Cocktail recipes are factual but may vary, small models may hallucinate variations
5	Determine if 37 is a prime number.	Medium	Simple math, but small models might struggle with number processing
6	What is the highest-rated movie of all time?	Hard	Ranking varies across sources; small LLMs may struggle to provide a definitive answer
7	Which actors are part of the film "A Love Song for Bobby Long"?	Hard	Niche movie knowledge; small LLMs might not have enough data on this film
8	Tell me what places are worth visiting in Tokyo.	Hard	Requires aggregation of recommendations; small models lack up-to-date travel data
9	Give me a recipe for pancakes without eggs.	Hard	Requires understanding of ingredient substitutions, which small LLMs struggle with

models for specialized knowledge areas. Future research can build upon these findings by refining evaluation methodologies and exploring techniques to enhance model reliability and efficiency, particularly in resource-constrained environments.

AUTHOR CONTRIBUTIONS

Research, Experiments, and Writing: V.P.-K., A.M., A.P., M.M., E.P., all of whom contributed equally. Mentorship and Guidance: P.R., B.I. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENT

This work is partially supported by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje

REFERENCES

[1] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, "A survey on large language models: Applications, challenges, limitations, and practical usage," *Authorea Preprints*, vol. 3, 2023.

[2] T. Xiao and J. Zhu, "Foundations of large language models," 2025. [Online]. Available: https://arxiv.org/abs/2501.09223

[3] S. Wang, S. Zhang, J. Zhang, R. Hu, X. Li, T. Zhang, J. Li, F. Wu, G. Wang, and E. Hovy, "Reinforcement learning enhanced llms: A survey," 2025. [Online]. Available: https://arxiv.org/abs/2412.10400

[4] G. Wang, S. Cheng, Q. Yu, and C. Liu, "Openchat: Advancing open-source language models with imperfect data," *version pre-release*, vol. 10, pp. 67–74, 2023.

[5] P. Zhang, G. Zeng, T. Wang, and W. Lu, "Tinyllama: An open-source small language model," *arXiv preprint arXiv:2401.02385*, 2024.

[6] M. J. Khan and G. Sukthankar, "Sc-phi2: A fine-tuned small language model for starcraft ii macromanagement tasks," *arXiv preprint arXiv:2409.18989*, 2024.

[7] R. Zhang, F. Song, H. Huang, J. Yuan, M. Yang, and T. Zhao, "Hit-mi&t lab's submission to eval4nlp 2023 shared task," in *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, 2023, pp. 139–148.

[8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: https://arxiv.org/abs/2307.09288

[9] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf, "Zephyr: Direct distillation of lm alignment," 2023. [Online]. Available: https://arxiv.org/abs/2310.16944

[10] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah, "Orca: Progressive learning from complex explanation traces of gpt-4," 2023. [Online]. Available: https://arxiv.org/abs/2306.02707

[11] Y. Albdeiwi and M. El-Khalil, "Generating test cases using natural language processing," *LU-CS-EX*, 2024.

[12] B. Zhu, E. Frick, T. Wu, H. Zhu, K. Ganesan, W.-L. Chiang, J. Zhang, and J. Jiao, "Starling-7b: Improving helpfulness and harmlessness with rlaif," in *First Conference on Language Modeling*, 2024.

[13] U. Iqbal, T. Kohno, and F. Roesner, "Llm platform security: Applying a systematic evaluation framework to openai's chatgpt plugins," 2024. [Online]. Available: https://arxiv.org/abs/2309.10254

[14] H. Lin, "Enhancing nlp robustness and generalization through llm-generated contrast sets: A scalable framework for systematic evaluation and adversarial training," 2025. [Online]. Available: https://arxiv.org/abs/2503.06648

[15] 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.

[16] B. Zirnstein, "Extended context for instructgpt with llamaindex," 2023.

Comparison of Hénon Map With Logistic Map Performance in Encryption and Decryption of Images

Zhivko Shishovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, North Macedonia
zhivko.shishovski@students.finki.ukim.mk

Aleksandar Stojmenski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, North Macedonia
aleksandar.stojmenski@finki.ukim.mk

Ivan Chorbev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, North Macedonia
ivan.chorbev@finki.ukim.mk

***Abstract*—Privacy, authentication, and confidentiality are very important during data transfer via various remote mediums such as the Internet. In this paper, we analyze two encryption algorithms and compare their performance. The chaotic cryptosystems known as Hénon map and Logistic map are compared by their properties, and two maps are tested on two data sources of medical and satellite images. The algorithms are evaluated by encryption and decryption speed values, the correlation coefficient between adjacent pixels, and output file size. The results showed that the Hénon map has an advantage in comparison with the Logistic map regarding performance.**

***Index Terms*—encryption, decryption, image, chaotic map, Hénon map, Logistic map.**

I. Introduction

The rapid progression of digital communication and network technologies has shown enormous potential in the storage and exchange of electronic data on the Internet. Medical images are important in diagnostics, so they are highly sensitive and specific and must be kept secure. In addition, multimedia shared and stored on the Internet is mostly in the form of images. There are multimedia data applied in military, government, industry and personal activities which are necessary to be protected before transmission [1]. Because of all that, network security has become a requirement to prevent data modification and intruders.

The confidentiality and authenticity of digital images are ensured by means of image encryption [2]. Any image encryption technique aims to convert the original image into a cipher image which is difficult to view. In the encryption research work there are three main tasks: the encrypted image not to show any information from the original image; run times of encryption and decryption processes has to be decreased; and the decrypted image has to be near equal to the original image [3].

Traditional ciphers function well for textual data, but they are not suitable for other types of multimedia data, which are usually bulky and very large-sized. They require a lot of time to directly encrypt multimedia for real-time operations and these algorithms also show significant latency [1], [3].

Numerous chaos-based image encryption techniques have been proposed [1], [2]. Permutation and diffusion are the two important stages involved in these techniques, which results in a higher level of security [4], [5]. Also, due to their high computational efficiency, high sensitivity to initial condition, system parameters, ergodicity properties, and pseudorandomness, they have advantages in use.

The aim of this study is the evaluation and comparison of two chaos-based encryption chippers: Hénon map and Logistic map. Their properties: dimensions, mathematical formulation, randomness, key generation efficiency, key space size, and security are compared. The two maps are tested at: encryption speed; decryption speed; correlation coefficient between adjacent pixels; and file sizes, on two different series of medical and satellite images.

II. Background Work

Chaos is a natural phenomenon discovered by E. Lorenz [6]. It is a movement created by a non-linear and ensured system. For a non-authorized user a chaotic signal looks like noise and ignores the mechanism for generating it [7]. If and only if the parameters are the same as the original value, the chaotic signal is reconstructed [7]. Important properties of chaotic systems include: lack of periodicity and topological transitivity, pseudo-random property, and high dependence on initial conditions and system parameters [4], [5], [7], [8].

The basic idea of using chaotic map for transformations data encryption is traced to the foundational papers of C. Shannon [9]. The most widely used chaos-based image encryption

scheme is proposed by J. Fridrich who adopted the Shannon approach for the design of secure cipher [10].

During recent years, various researchers investigated chaos-based image encryption [11]. Most of them are based on two principle processes: confusion and diffusion. In each round of the encryption process, pixel positions are initially scrambled in a hidden way, reducing the similarity between neighboring pixels. After that, the pixel values are altered chronologically, influencing the modification process of each pixel and the diffusion of all its subsequent pixels. A completely new cipher image is created by performing multiple rounds of encryption with a small modification of the plain image [7].

III. Experimental Work

Experimental setup

In order to analyze both encryption maps, we used the Python programming language along with its Open-CV library (named "cv2") to read and write images and the numpy library to simplify the execution of mathematical operations. The codes are executed on a computer with a processor with a base speed of 3.20 GHz.

Data Sources

The experiments were carried out on data sets of images from the medical domain and images from geographical satellites. The medical source consisted of 100 EchoNet-Dynamic [12] samples with lossless PNG compression quality of size 112 x 112 pixels with similar content, the second set consisted of 7 Google Maps satellite images with lossy JPEG compression quality of size 256 x 256 pixels with distinguishable content [13].

Hénon Map Encryption

The Hénon map [14], [15], [17] is a chaotic system defined as:

$$x_{n+1} = 1 - ax_n^2 + y_n \tag{1}$$

$$y_{n+1} = bx_n \tag{2}$$

where:

- x_n and y_n are the sequence values,
- a and b are control parameters (commonly $a = 1.4$, $b = 0.3$),
- x_0 and y_0 are the initial conditions.

Encryption Formula

Each pixel is encrypted using a chaotic key stream derived from the Hénon map:

$$C_{i,j,k} = P_{i,j,k} \oplus \lfloor (x_{i,j} - \min(x)) \times \frac{256}{\max(x) - \min(x)} \rfloor \tag{3}$$

Decryption Formula

The decryption process follows the same XOR operation:

$$P_{i,j,k} = C_{i,j,k} \oplus \lfloor (x_{i,j} - \min(x)) \times \frac{256}{\max(x) - \min(x)} \rfloor \tag{4}$$

Algorithm 1 Hénon map iteration

Require: x_0, y_0, a, b, N ▷ Initial values, parameters, and iterations
Ensure: Sequence of (x_n, y_n) values
1: $x \leftarrow x_0$
2: $y \leftarrow y_0$
3: **for** $n = 1$ to N **do**
4: $x_{\text{new}} \leftarrow 1 - ax^2 + y$
5: $y_{\text{new}} \leftarrow bx$
6: $x \leftarrow x_{\text{new}}$
7: $y \leftarrow y_{\text{new}}$
8: **Output** (x, y)
9: **end for**

Logistic map encryption

The Logistic map [16], [17] is defined as:

$$x_{n+1} = rx_n(1 - x_n) \tag{5}$$

where:

- x_n is the sequence value at step n,
- r is the control parameter (commonly 3.99 for high chaos),
- x_0 is the initial seed ($0 < x_0 < 1$).

Encryption Formula

Each pixel in the input image (P) is encrypted using the Logistic map as:

$$C_{i,j,k} = P_{i,j,k} \oplus \lfloor x_{i,j} \times 256 \rfloor \tag{6}$$

Decryption Formula

Since XOR is a reversible operation, applying it again recovers the original pixel (P):

$$P_{i,j,k} = C_{i,j,k} \oplus \lfloor x_{i,j} \times 256 \rfloor \tag{7}$$

IV. Discussion

Firstly, the Hénon map and Logistic map were compared by their properties, and further, their encryption and decryption efficiencies were tested on two series of images: medical and satellite.

Hénon is two-dimensional map (2D), it has two variables and can provide more complex encryption than Logistic map which is one-dimensional (1D) and has single variable.

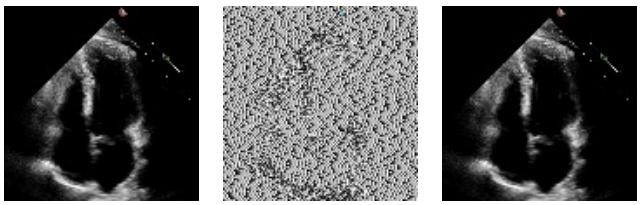
The Hénon map, defined by the equations (1) and (2), and the Logistic map is defined by the relation (5). According to their mathematical formulation, both maps display elementary dynamics and are characterized by computational efficiency.

Both, the Hénon map and the Logistic map, due to their chaotic behavior, generate random key sequences.

The Hénon map is computationally efficient for key generation, due to its simple 2D equations and the Logistic map presents a computationally efficient approach for key generation owing to its 1D nature and direct iteration formula.

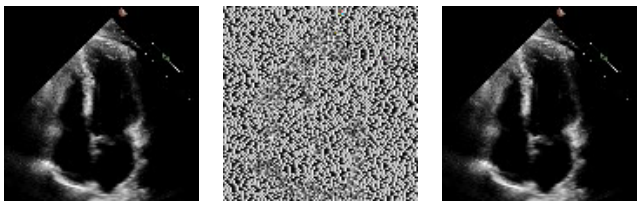
The Hénon mapping exhibits a 2D domain of keys, which presents a superior key space (2^{64}) in contrast to the key space of the Logistic map (2^{32}) [17].

Measuring the computational efficiency of the key-map considers the time taken to generate key sequences. Faster maps can lead to quicker encryption and decryption. Figures from Fig. 1 to Fig. 4 show the image encryption-decryption procedure results given for the two sets of images, for the two chaotic maps used (Hénon map and Logistic map).



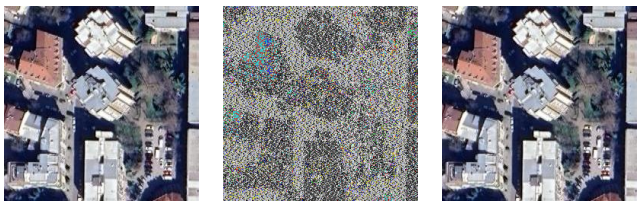
Original Image Encrypted Image Decrypted Image

Fig. 1: Hénon map results for medical images



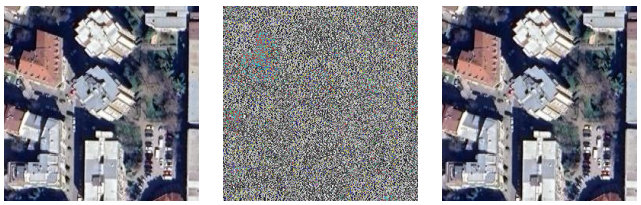
Original Image Encrypted Image Decrypted Image

Fig. 2: Logistic map results for medical images



Original Image Encrypted Image Decrypted Image

Fig. 3: Hénon map results for satellite images



Original Image Encrypted Image Decrypted Image

Fig. 4: Logistic map results for satellite images

The results obtained for encryption and decryption speed of Hénon map and Logistic map used on medical and satellite images are shown at Fig. 5 and Table. I.

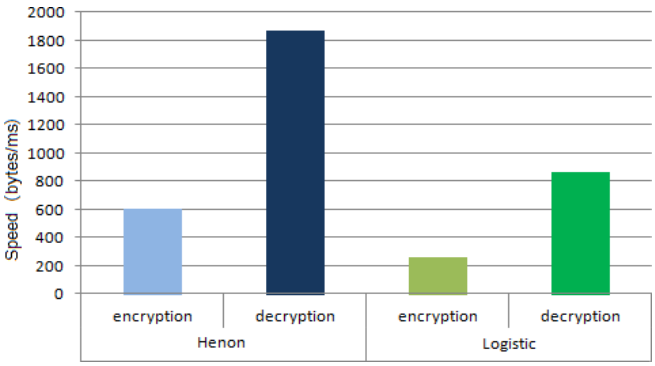


Fig. 5: Encryption and decryption speed using Hénon and Logistic map.

TABLE I: Hénon map and Logistic map encryption and decryption speed

Average	Hénon map		Logistic map	
File Size (bytes)	Encryption (bytes/ms)	Decryption (bytes/ms)	Encryption (bytes/ms)	Decryption (bytes/ms)
[a]11266	608.14	1870.44	263.61	860.80
[b]14561	162.46	2132.39	70.64	962.11

[a]Results are obtained by processing 100 similar medical images.
[b]Results are obtained by processing 7 various satellite images.

These results prove that the Hénon map has superiority over Logistic map at encryption and decryption speed.

File sizes of images compared with original image size are presented at Fig. 6.

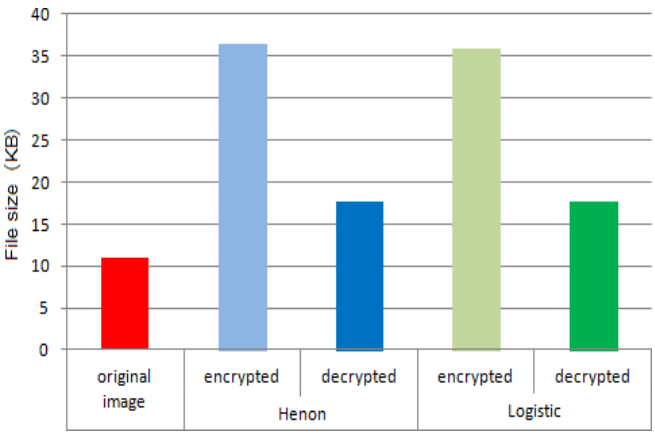
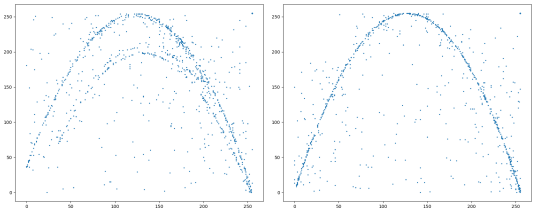


Fig. 6: File sizes before and after encryption and decryption medical images using Hénon and Logistic map.

The results obtained for file sizes show that both maps generate files with similar sizes.

Adjacent pixel correlation test results are given at Fig. 7.



Hénon Logistic

Fig. 7: Adjacent pixel correlation comparisons

The adjacent pixel correlation test results show that the correlation coefficients of the encrypted map images are near zero and the two encryption systems make the adjacent pixels very independent.

Chaotic key maps exhibit high sensitivity to initial conditions and minor changes in initial values which result in distinct key sequences. Additionally, they possess resistance against cryptographic attacks. A larger chaotic key space enhances security [17]. This means that Hénon map has advantage in comparison with Logistic map.

Conclusion

The Hénon map and the Logistic map comparison and testing on two different series of medical and satellite images had shown the following:

The Hénon map has advantages in dimension (2D), in mathematical formulation (two dimensional formulas); in key space size (2^{64}); in security and in computational efficiency. Testing the encryption and decryption speed, it shows this map's superiority (for approximately 2.2 times). This map is chaotic, shows dynamics and randomness. The Hénon map offers a balance between computational efficiency and security, it exhibits suitability for implementation in lightweight encryption scenarios and image encryption endeavors owing to its dynamic properties in two dimensions.

The Logistic map is simpler, one-dimensional (1D), although weaker in mathematical formulation (one-dimensional formula), it has smaller key space size (2^{32}). Also, it is weaker in security and in computational efficiency. Testing the encryption and decryption speed shows this map's disadvantage in correlation with Hénon map. This map is also chaotic, shows dynamics and randomness. The Logistic map is a fitting choice for straightforward encryption techniques and stream ciphers owing to its efficiency and susceptibility to initial conditions. Lower space complexity is preferable for efficient implementation.

Similar file sizes are obtained using the both maps, during the encryption and the decryption processes. The results obtained using the adjacent pixel correlation test show strong de-correlation, which makes a statistical attack on these two cryptosystems difficult.

Chaotic encryption algorithms are the preferred choice when it comes to reliable security during multimedia transfer. The Hénon map provides faster execution of the encryption and decryption process of images than the Logistic map, but has a weakness of showing the original image's shape patterns (Fig. 3).

For future studies, other approaches might be utilized such as Lorenz map with the usage of a 3D map which would enable to compare its performance, and to investigate its advantages and limitations. Furthermore, the research could be extended by additional test parameters and suitable applications of chaotic maps.

References

[1] N. Mekki, M. Hamdi, T. Aguili and T. Kim, "A Real-time chaotic encryption for multimedia data and application to secure surveillance framework for IoT system", Conference Paper, April 2018. DOI: 10.1109/COMMNET.2018.8360271

[2] U. Zia, M. McCartney, B. Scotney, J. Martinez, M. AbuTair, J. Memon and A. Sajjad, "Survey on image encryption techniques using chaotic maps in spatial, transform and spatiotemporal domains", International Journal of Information Security, 2022, vol. 21, pp. 917–935. [Online]. Available: https://doi.org/10.1007/s10207-022-00588-5

[3] M. T. Saleh, M. A. W. Shalaby and H. N. Elmahdy, "Modified Arnold's cat map-RC4 encryption technique for medical images", Egyptian Computer Science Journal, vol. 44, no. 2, May 2020. ISSN-1110-2586

[4] Z.-H. Guan, F. Huang and W. Guan, "Chaos-Based image encryption algorithm", Physics Letters A, vol. 346, no. 1-3, pp. 153-157, October 2005.

[5] R. Kumar, A. Sampath and P. Indumathi, "Enhancement and analysis of chaotic image encryption algorithms", Conference Paper, (CCSEA 2011), 2011. DOI: 10.5121/csit.2011.1215

[6] E. N. Lorenz, "Deterministic nonperiodic flow", Journal of Atmospheric Sciences, vol. 20, no. 2, pp. 130-141, 1963.

[7] S. S. Hassan, Z. Iqbal, M. J. Ikram and M. Ishaque, "Hénon chaotic map based image encryption scheme using bit-level circular shift", Journal of Theoretical and Applied Information Technology, vol.100, no. 6, March 2022. ISSN: 1992-864, E-ISSN: 1817-3195

[8] Y. Mao, G. Chen and S. Lian, "A Novel fast image encryption scheme based on 3D chaotic baker maps", International Journal of Bifurcation and Chaos, vol. 14, no. 10, October 2004. DOI: 10.1142/S021812740401151X

[9] C. E. Shannon, "Communication theory of secrecy systems", Bell System Technical Journal, vol. 28, no. 4, pp. 656-715, 1949.

[10] J. Fridrich, "Symmetric ciphers based on two-dimensional chaotic maps", International Journal of Bifurcation and Chaos, vol. 8, no. 6, pp. 1259–1284, 1998.

[11] M. Yildirim, "Comparative analysis of the effects of chaotic systems on the robustness of image encryption", Electrica, vol. 21, no. 2, pp. 209–215, May 2021. DOI: 10.5152/electrica.2021.21027

[12] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley and J.Y. Zou, "Video-based AI for beat-to-beat assessment of cardiac function", Nature, vol. 580, no. 7802, pp. 252-256, April 2020. DOI: 10.1038/s41586-020-2145-8.

[13] H. Humenberger, "Explorations and estimations using Google Maps images", International Journal of Mathematical Education in Science and Technology, vol. 53, no. 2, pp. 503-515, 2022. DOI: 10.1080/0020739X.2020.1863489

[14] F. Qi, S. Huang, T. Li, H. Yang and X. Kang, "2D Hénon-Chebyshev chaotic map for image encryption", Conference paper, 21st International Conference on High and Performance Computing and Communications (IEEE), August 2019. DOI: 10.1109/HPCC/SmartCity/DSS.2019.00114

[15] M. Sharma and A. Sharma, "A Secret file sharing scheme with chaos based encryption", Conference paper, 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), July 2019. DOI: 10.1109/ICCCNT45670.2019.8944344

[16] Y. Dai and X. Wang, "Medical image encryption based on a composition of Logistic maps and Chebyshev maps", Conference paper, International Conference on Information and Automation Shenyang (IEEE), China, June 2012. DOI: 10.1109/ICInfA.2012.6246810

[17] D. S. Salman and J. R. Naif, "Comparative study of chaotic system for encryption", Iraqi Journal for Computers and Informatics, vol. 49, no. 2, pp. 83-92, 2023.

AI-Driven Project Management: LLM-Powered Task Automation and Optimization

1st Ilija Bozoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, North Macedonia
ilija.bozoski@students.finki.ukim.mk

2nd Sasho Gramatikov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, North Macedonia
sasho.gramatikov@finki.ukim.mk

***Abstract*—The integration of artificial intelligence (AI) in project management has the potential to automate routine tasks, optimize user assignment, and improve decision-making. This paper presents an AI-enhanced project management system built in Laravel, incorporating automated task generation using Llama 3.2 3B, AI-driven user recommendations, task classification, and a Neo4j knowledge graph for efficient data management. The system significantly improves project efficiency by automating key processes and enhancing collaboration. The experimental results demonstrate improved task categorization accuracy and optimized user task assignments.**

***Index Terms*—Artificial Intelligence (AI), Project Management, Laravel Framework, Task Automation, Large Language Models (LLMs), User Recommendation Systems, Neo4j Knowledge Graph, Task Classification, LLaMA 3.2 3B**

I. Introduction

Efficient task management is crucial for project success, yet traditional systems often rely on manual allocation, leading to inefficiencies like resource mismanagement [1]. Modern project managers must combine technical expertise with skills in data analysis and stakeholder management to leverage AI-driven advancements [2]. While tools like Jira provide structured workflows, they still require extensive manual configuration. AI integration revolutionizes this process by automating task creation, intelligent assignment, and progress tracking, improving both productivity and adaptability.

Unlike prior work—such as [3], which uses GPT-3.5 for basic task automation—our system employs the advanced LLaMA 3.2-3B model, an open-source alternative that offers superior flexibility, customization options, and operational transparency. This innovative approach not only addresses the critical standardization question raised in [4] for AI in project management but also provides measurable improvements in deployment efficiency. While existing platforms (e.g., Jira) use AI reactively for simple suggestions, our system actively shapes project execution through intelligent task generation, context-aware prioritization of workflows, and optimal resource assignments based on real-time team capacity analysis. Prior research [5][6] emphasizes AI's supporting role in risk detection while maintaining human oversight; our approach represents a paradigm shift by enabling secure autonomous task delegation and continuous workflow optimization, significantly reducing both cognitive bias and traditional managerial overhead.

By integrating real-time AI insights, the system enables proactive bottleneck detection and evidence-based adjustments, surpassing conventional tools. It transforms project management into an adaptive, self-regulating framework that improves team communication and outcomes.

The remainder details system architecture (Section 2), evaluation results (3.1), and future advancements (3.2).

II. System Architecture and Methodology

A. System Components

The proposed system consists of multiple interconnected components, each playing a critical role in ensuring efficient project management. All components are securely connected within a private network, ensuring controlled access and enhanced security. The primary components are:

- Laravel Application: The Laravel framework [7] serves as the core back-end system, providing a robust foundation for managing entities such as companies, projects, users, tasks, comments, and notifications. In addition to these basic functionalities, the application is responsible for critical system operations, including authentication, authorization, and task rating. It also implements a personalized email notification system, which triggers emails for events such as new user registration, task creation, and task status updates. Furthermore, the application exposes a set of APIs to facilitate user interactions and seamlessly orchestrates communication with external AI services, ensuring a cohesive and integrated workflow.
- Nginx Web Server: Nginx acts as a reverse proxy, efficiently managing incoming requests to the Laravel application. It enhances performance through load balancing and caching while improving security by filtering and managing traffic before it reaches the application server.
- PostgreSQL Database: The system employs a PostgreSQL database to store all structured data, including companies, projects, users, tasks, task categories, user positions, and comments. It ensures data persistence, relational consistency, and efficient querying through normalized table structures. Beyond its traditional role,

the database is also utilized as a job queue for scheduled tasks, such as sending notifications and emails. This dual-purpose design simplifies the architecture by leveraging the database's reliability and transactional capabilities for both data storage and background job management.

- FastAPI Backend: lightweight, independent service built with FastAPI facilitates AI-driven functionalities by acting as a bridge between the Laravel application and external AI services. This service periodically retrieves entities (e.g., users, tasks, projects) from Laravel via API requests, processes the data, and constructs a knowledge graph in Neo4j. The process involves creating nodes for each entity and establishing relationships between them, such as user-task assignments or project-task dependencies. Scheduled to run every four days, the entire process completes in approximately one minute, ensuring the knowledge graph remains up-to-date and reflective of the latest project state.
- Neo4j Knowledge Graph Database: Neo4j[8] serves as the system's graph database, storing relationships between entities such as projects, tasks, and users. This enables efficient retrieval of structured knowledge, supporting advanced querying and decision-making. Beyond its role as a database, Neo4j provides a powerful graph visualization tool through its built-in browser interface. This tool allows users to interactively query and explore all nodes and relationships, offering insights into complex dependencies and patterns within the data. Together, these features make Neo4j a central component for both data storage and knowledge discovery in the system.
- LLaMA Model Server: A dedicated server hosts the LLaMA 3.2 3B model [9], providing AI-driven functionalities such as automatic task generation, task classification, task-specific chatbot assistance, and user recommendation. Accessed via HTTP endpoints, the server processes prompts sent from both the Laravel and FastAPI backends. For example, it generates structured tasks based on project descriptions, classifies tasks into categories, provides context-aware chatbot responses for task-related queries, and suggests the most suitable users for specific tasks based on their skills and availability. This centralized AI server enhances the system's intelligence and automation capabilities.

B. AI-Powered Features

While these features significantly enhance the system's capabilities, there are instances where the language model fails to generate a valid response or produces outputs in an unexpected format. To address such cases, fallback mechanisms have been implemented. A common approach involves re-sending the same prompt multiple times (typically up to three attempts) to increase the likelihood of obtaining a valid result. If all attempts fail, the system gracefully bypasses the AI-enhanced processes and continues to function in a standard, non-AI mode to ensure operational continuity.

1) Automated Task Creation: The system provides two modes for project creation:AI-powered and regular. If the AI-powered option is selected, the automated task generation process is triggered immediately after project creation. This process is facilitated by a request sent to the LLaMA model server, containing the project description and predefined task categories. Upon receiving the request, the large language model (LLM) analyzes the project description and performs the following operations:

- Task Generation: The model decomposes the project into structured tasks.
- Priority Assignment: Each task is assigned a priority level (*low, medium, or high*) based on its urgency and importance.
- Task Classification: The tasks are categorized into predefined categories

Once the LLaMA server returns the generated tasks, they are processed through a *task processing pipeline*, where relevant features such as *title*, *description*, *priority*, and *category* are extracted. The Laravel backend then creates and stores the tasks in the PostgreSQL database, associating them with the corresponding project.

This automated pipeline improves efficiency by reducing manual effort in structuring projects and ensuring that task allocation aligns with project objectives.

2) Task Classification: In the proposed system, tasks generated by a LLLM are categorized as part of the task generation process, based on their title and description. This step ensures task relevance and facilitates user-task matching. A predefined set of approximately 15 categories is provided, and the LLM assigns each new task to one of these categories, enhancing efficiency and suitability during task creation.

3) AI-Driven Task Assistance Chatbot: To assist users who encounter difficulties while working on a task, a chatbot is implemented as part of the system. When a user is stuck, they simply need to input their question. In the background, both the task description and the user's question are sent to the LL server. The LLM generates detailed and descriptive responses, often supplemented with examples, to provide clear and actionable guidance to the user. This feature ensures effective support and enhances the user's ability to complete tasks efficiently.

4) Most Suitable User Recommendation: To optimize task assignment, the system employs a hybrid AI-human framework that combines performance metrics with LLM-based analysis. Upon task completion, each user receives a composite rating:

- **Manager's Rating**, reflecting task quality and collaboration on previously assigned tasks.
- **System Rating**, evaluating efficiency, response time, and deadline adherence.

The system continuously tracks user-specific metrics, including category-specific average ratings, total completed tasks, and current workload (unfinished tasks). When a new task is created, 15–20 candidates are shortlisted based on these indicators.

The LLM evaluates the shortlisted users by analyzing position relevance, expertise, past performance, and current workload. Notably, users with a high number of open tasks are deprioritized, even if they have excellent ratings, to ensure balanced task distribution and maintain efficiency.

The top four users are recommended to the task manager, who makes the final assignment decision, preserving human oversight.

5) Knowledge Graph Implementation: The system uses a knowledge graph to model key entities and their relationships, focusing on the essential components to understand the functionality of the system. This graph is implemented using Neo4j and integrated via a FastAPI service. Every four days, the FastAPI backend collects relevant data from the system and updates the Neo4j database by creating nodes and relationships. The Neo4j browser serves as an interactive interface for visualizing the graph structure, including models, relationships, and system interactions. The system administrator has direct access to the Neo4j browser, allowing for intuitive exploration of the project structures and dependencies through interactive queries. This implementation improves system transparency, providing a clear and dynamic representation of the project management processes.

III. Results and Evaluation

A. System Performance and Accuracy

The system's performance was assessed by evaluating the efficiency of automated task generation, the accuracy of task classification, the quality of assistance provided by the chatbot, and the effectiveness of user recommendations.

1) Automated Task Creation: The automated task creation process was evaluated on the basis of efficiency, time reduction, and task quality. The results demonstrate a significant improvement compared to manual task creation. Traditionally, project setup, including the creation of eight structured and categorized tasks, requires approximately 10 to 15 minutes, even when tasks are predefined. However, utilizing the AI-powered project creation system, this process is reduced to approximately 1.5 to 2.5 minutes, marking a substantial efficiency improvement.

2) Task Classification Accuracy: The evaluation of task classification accuracy was conducted using two distinct approaches.

In the first approach, the focus was on the correctness of AI-assigned categories in 128 classified tasks. The analysis revealed that 8 tasks were misclassified, resulting in an accuracy rate of approximately 93.75

The second approach involved a comparative analysis between DeepSeek's classification output and ChatGPT's categorization, both benchmarked against a reference classification generated by a LLaMA-based model of the system. The results indicate that DeepSeek's classification closely aligns with the LLaMA model, achieving an accuracy exceeding 98%.

In contrast, ChatGPT exhibited a divergent categorization pattern. While ChatGPT's classifications differed from those of the LLaMA model, it notably confined all assigned categories to only four distinct classes, with a substantial portion of tasks labeled as "Undefined". This suggests a tendency towards under-specification in ambiguous or uncertain task descriptions.

3) User Recommendation Effectiveness: The introduction of an AI-assisted user selection mechanism has significantly optimized the task assignment process within the system. Traditionally, managers were required to manually select a suitable user from a pool of approximately 60 candidates. This manual process was not only time-consuming but also prone to subjectivity and inconsistency.

In contrast, the AI-enhanced approach narrows down the selection to the 4 most suitable users based on task requirements and user profiles. This has resulted in substantial time savings and has improved overall selection efficiency.

To benchmark performance, we compared the user recommendations generated by ChatGPT, DeepSeek, and the proprietary system, all using the same input prompt for identical tasks. The results demonstrated notable overlap in some cases, while also revealing unique selections based on the models' differing evaluation strategies (see Figure 1).

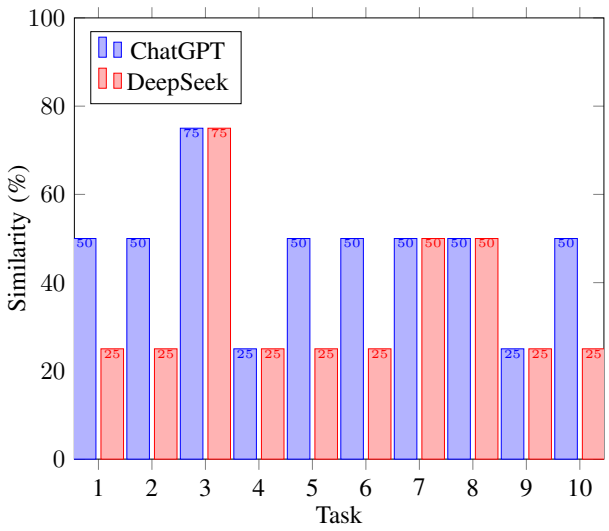


Fig. 1. Similarity of number of members suggested by the proposed system, ChatGPT and DeepSeek.

This streamlined selection process enhances productivity and consistency, while reducing managerial overhead during task delegation.

4) AI-Driven Task Assistance Chatbot: The AI-driven task assistance chatbot provides detailed explanations for each specified task, responding effectively to user inquiries in most cases. Compared to widely used models such as ChatGPT and DeepSeek, the chatbot generates responses that are longer, more structured, and developer-oriented. This makes it particularly useful for technical guidance and project-related queries.

A key strength of the chatbot is its ability to generate well-structured code snippets and provide technology recommen-

dations. For instance, when comparing front-end frameworks such as React and Angular, the chatbot delivers insights into their advantages and best use cases. This feature enhances the decision-making process for developers.

However, a limitation of the model is that its responses may sometimes be based on outdated information, as it was trained on earlier datasets. Despite this, the chatbot remains highly effective in assisting users with understanding task requirements, suggesting optimal approaches, and improving overall task efficiency.

5) Knowledge Graph: The knowledge graph within the system is fully rebuilt on a regular cycle, specifically every four days. The rebuilding process itself is highly efficient, with an average execution time of approximately 35 to 40 seconds. The implementation of the knowledge graph has proven to be an invaluable tool in improving the understanding and familiarity of the system entities.The knowledge graph enhances system understanding by visually representing key relationships between entities. Figure 2 shows only connected nodes, with users (blue), projects (orange), positions (red), tasks (light brown), task categories (green), and companies (purple), along with relationships like Manager → project manager and Works_on → user assigned to a task. Some users, such as the admin, appear more connected due to system testing, but this reflects how the graph would evolve over 8–10 months, and specific queries can generate more refined insights.structures.

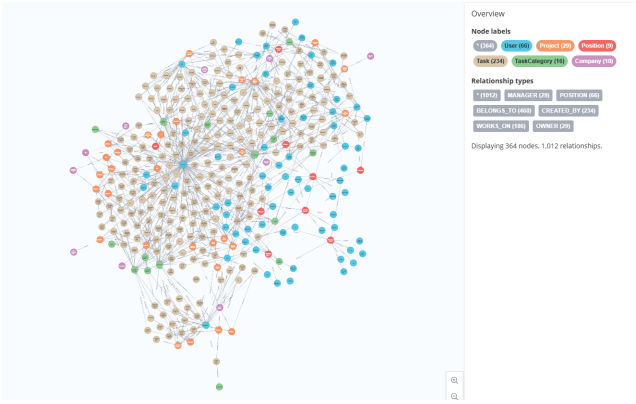


Fig. 2. Knowledge graph visualization highlighting core processes and workload distribution.

This visual representation further assists in pinpointing operational bottlenecks and inefficiencies, thereby contributing to improved system performance and workforce management.

B. Discusssion

The AI components within this system are primarily based on the LLaMA 3.2 3B model, which, although effective, has inherent limitations due to its relatively small size. This constraint affects the model's ability to generate highly sophisticated responses and perform complex reasoning tasks. Furthermore, the training data on which the model was built is slightly outdated, which could affect the accuracy and relevance of chatbot responses.

The system is fully containerized using Docker, with the exception of the external server hosting the large language model (LLM). This architecture allows for easy deployment across different environments and supports future scalability through tools like Kubernetes or cloud platforms. However, a key limitation lies in the computational infrastructure, as the system was developed and deployed on a personal computer. This constrains processing power and currently limits the use of larger models with more advanced reasoning capabilities.

IV. Conclusion

In this paper, an AI-driven solution for project management is proposed, integrating automation for task generation, user assignment, and progress monitoring. The system is designed to enhance adaptability while maintaining efficiency and usability. The evaluation results demonstrate that it significantly improves project's manager efficiency by reducing manual workload and streamlining task distribution. Notably, even a relatively compact model has proven capable of handling complex processes—such as intelligent user suggestions—demonstrating that strategic optimization can deliver high performance without excessive computational demands. Future work will focus on enhancing accessibility through speech-to-text integration, leveraging large language models (LLMs) to refine predictive analytics and task automation, and expanding real-world testing across diverse industries. These efforts will not only validate the system's scalability but also uncover actionable insights for further optimization, ensuring it meets the evolving needs of modern project teams.

References

[1] R. Bhanderi, "AI-Driven Project Management Revolutionizing Workflow Optimization and Decision-Making," International Journal of Trend in Scientific Research and Development (IJTSRD), vol. 8, no. 6, pp. 325–338, Dec. 2024, doi: 10.xxxx/ijtsrd71577.

[2] M. J. Karamthulla, A. Tadimarri, R. Tillu, and M. Muthusubramanian, "Navigating the Future: AI-Driven Project Management in the Digital Era," in IJFMR ,vol 6 i 2,pp. XX-XX, March-April 2024,

[3] M. Jayaram, Y. Bhutkar, I. L. K. Bojjanapalli, G. Yeshwanth, and B. Y. Reddy, "Beyond Automation: AI-Driven Project Management with OpenAI and Prompt Engineering," in 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET), Sydney, Australia, Jul. 25–27, 2024, pp. [Page Numbers], doi: 10.1109/ICECET61485.2024.10698333.

[4] M. H. Salleh and K. A. Aziz, "Artificial Intelligence Augmented Project Management," in *International Conference on Technology and Innovation Management (ICTIM 2022)*, Atlantis Press, December 2022, pp. 274–284.

[5] Md. Mahfuzul Islam Shamim, "Artificial Intelligence in Project Management: Enhancing Efficiency and Decision-Making," *International Journal of Management Information Systems and Data Science*, vol. 1, no. 1, pp. XX-XX, April 2024.

[6] Shadevan, S. (2023). Project Management in the Era of Artificial Intelligence. *European Journal of Theoretical and Applied Sciences*, 1(3), 349–359. DOI: 10.59324/ejtas.2023.1(3).35.

[7] Laravel, *Laravel Documentation (v11)*. 2024. [Online]. Available: https://laravel.com/docs/11.x

[8] Neo4j. (2024). "Cypher Query Language Guide." *Neo4j Documentation*. [Online]. Available: https://neo4j.com/docs/cypher-manual/current/

[9] Meta AI, "Meta-Llama-3-3B," Hugging Face, 2024. [Online]. Available: https://huggingface.co/meta-llama/Meta-Llama-3-3B

Data Preprocessing and Exploration of Developmental Disabilities Dataset

1st Ana Markovska
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
ana.markovska.3@students.finki.ukim.mk

2nd Bojan Ilijoski, PhD
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
bojan.ilijoski@finki.ukim.mk

***Abstract*—This study focuses on the exploratory data analysis (EDA) and preprocessing of a dataset collected as part of the Erasmus+ SynergyEd project. The dataset, originally published in BMC Pediatrics under the title "Perceptions of support, challenges, and needs among parents/caregivers of children with developmental disabilities in Croatia, North Macedonia, and Serbia", captures parental perceptions and experiences related to children with developmental disabilities in Croatia, North Macedonia, and Serbia. The dataset consists of 953 survey responses, covering a range of variables related to child characteristics, service encounters, caregiver perceptions, and demographic factors. Data preprocessing involved standardization, normalization, missing data handling, and categorization of diagnoses. Several machine learning techniques, including clustering algorithms, were explored for grouping diagnostic information, but manual classification using regular expressions proved to be the most effective approach. This paper provides a detailed account of the data preprocessing steps, challenges encountered, and key insights from EDA, which lay the foundation for further statistical modeling and analysis.**

***Index Terms*—dataset, developmental disabilities, exploratory data analysis, data preprocessing**

I. Introduction

Children with developmental disabilities face various challenges that impact their cognitive, social, and physical well-being. Families of these children often encounter difficulties in accessing services, receiving appropriate support, and understanding their child's condition. To better comprehend these challenges, a cross-sectional survey was conducted as part of the Erasmus+ SynergyEd project. The collected dataset provides a unique opportunity to analyze trends in parental experiences, service accessibility, and child development across three countries: Croatia, North Macedonia, and Serbia [1].

The dataset consists of responses from 953 parents and caregivers, with the highest representation from Croatia (542 participants), followed by Serbia (206 participants) and North Macedonia (205 participants) [2]. This diverse sample allows for meaningful cross-country comparisons and a broader understanding of regional trends. The survey, conducted between March and April 2023, captures a snapshot of the current landscape of developmental disability support systems in these countries.

In addition to quantitative data on diagnoses, services, and daily routines, the dataset captures subjective caregiver experiences, such as emotional burden, perceived discrimination, and satisfaction with available support. These qualitative variables provide deeper insight into the lived realities and unmet needs of families raising children with developmental disabilities.

To ensure data consistency and facilitate analysis, extensive preprocessing was required. The dataset was cleaned, standardized, and consolidated before exploratory data analysis (EDA) was performed to understand its structure and key trends. Several machine learning experiments were conducted to explore patterns in diagnostic classifications, but traditional automated clustering methods failed to produce meaningful groupings. Instead, a manual regular expression (regex) mapping approach was implemented to categorize diagnoses effectively. This paper provides a detailed account of these preprocessing and exploratory data analysis steps, highlighting key challenges and insights.

II. Related work

Exploratory Data Analysis (EDA) and data preprocessing are essential steps in preparing datasets for meaningful insights, particularly in healthcare and social science research. Several studies have explored these methodologies, emphasizing their role in improving data quality, uncovering patterns, and facilitating further analysis.

EDA is widely used in healthcare data analysis to identify trends, detect anomalies, and enhance decision-making. Researchers have highlighted the importance of data visualization, summary statistics, and correlation analysis in extracting valuable insights from healthcare datasets [3]. Python-based EDA techniques leveraging libraries such as Pandas, NumPy, and Matplotlib have also been employed to clean data, manage missing values, and transform variables for predictive modeling and evidence-based decision-making [4].

Beyond healthcare, EDA has been applied in various domains to discover hidden patterns and relationships in datasets. In the service sector, EDA has been shown to facilitate anomaly detection, hypothesis testing, and data validation before applying machine learning models [5]. These techniques are essential in handling complex datasets, ensuring that data-driven insights are reliable and actionable.

Preprocessing techniques play a critical role in data analysis, particularly when dealing with noisy, incomplete, or hetero-

geneous datasets. Studies have demonstrated that selecting appropriate preprocessing methods—such as normalization, missing data handling, and attribute selection—can significantly impact the effectiveness of classification models in medical data analysis [6]. A structured approach to EDA has also been proposed for clinical datasets, incorporating outlier detection, clustering algorithms, and Principal Component Analysis to identify meaningful patient subgroups [7]. These methodologies align closely with the preprocessing techniques used in this study, where diagnostic classification was enhanced using regular expressions and literature-based mapping of symptoms.

Furthermore, EDA has been effectively used in healthcare classification tasks, helping researchers analyze demographic distributions, disease prevalence, and healthcare utilization patterns [8]. Such approaches provide a foundation for understanding datasets related to developmental disabilities, as they enable the identification of key characteristics and trends in patient and caregiver experiences.

These studies collectively highlight the importance of EDA and preprocessing in ensuring data integrity, uncovering hidden insights, and improving analytical outcomes. The methodologies explored in prior research serve as a basis for this study's approach to handling and analyzing data on parental perceptions of children with developmental disabilities.

III. Data Preprocessing

A. Data Integration and Standardization

The initial dataset was divided into three separate sheets representing survey responses from Croatia, North Macedonia, and Serbia. To enable comprehensive analysis, these sheets were merged into a single dataset. During the merging process, redundant columns were identified and combined to maintain consistency.

The original dataset contained column names and variable labels in multiple languages, with some values recorded in local languages and others in English. To standardize the dataset and improve readability, all column headers and response values were translated into English. This ensured consistency and facilitated easier interpretation during subsequent analysis.

To support country-specific comparisons, a new column labeled "country" was introduced, identifying the country of origin for each response. This variable allowed for filtering and grouping data by country, enhancing the analysis of regional trends.

B. Diagnosis Standardization

One of the most significant challenges during preprocessing was the variability in diagnostic labels. Diagnoses were presented in both Latin and Cyrillic scripts, leading to inconsistencies that needed to be resolved for accurate analysis. To address this issue, a new column titled "Child's Current Diagnosis (Latin)" was introduced, where all Cyrillic entries were systematically converted into Latin script. This step ensured that all diagnoses were represented in a single script, improving readability and consistency across the dataset.

However, script conversion alone was not sufficient to achieve full uniformity. Spelling variations, typos, and differences in formatting still posed challenges. To mitigate these inconsistencies, an additional column titled "Normalized Diagnosis" was created. In this column, all diagnoses were converted to lowercase letters to eliminate case sensitivity issues. Furthermore, regular expressions (regex) were employed to standardize variations in spelling and phrasing, ensuring that similar diagnoses were grouped correctly.

To enhance clarity and facilitate meaningful analysis, each diagnosis was mapped to one of 11 predefined diagnosis groups: Autism Spectrum (F84), Speech and Language Disorders (F80), Developmental Delay (R62), Intellectual Disability, ADHD (F90.0), Epilepsy, Cerebral Palsy (G80.9), Down Syndrome (Q90.9), Chromosomal Disorders, Neurological Disorders and Mixed Developmental Disorder (F83).

Diagnoses that did not fit within these categories were assigned to the "Other or Unclassified" category to ensure no data was omitted. This structured classification process not only improved the uniformity of the dataset but also enabled clearer insights into diagnosis distribution and patterns.

IV. Handling Missing Values

The dataset initially contained 954 responses; however, 224 entries were completely blank and subsequently removed to ensure the dataset included only valid responses. This step reduced the dataset to 730 valid entries. Addressing partially missing data required careful consideration. In particular, some survey questions — especially those involving agreement or satisfaction scales — were left unanswered. Since non-responses could either indicate disagreement or an intentional decision not to answer, a logical imputation strategy was applied. If a parent had responded affirmatively to adjacent questions but left one response blank, it was assumed that the missing response indicated disagreement. Conversely, if multiple consecutive questions were left unanswered, the missing responses were treated as non-responses rather than disagreement. This strategy minimized bias while ensuring the dataset accurately reflected participants' perceptions.

V. Machine Learning Experiments for Diagnosis Categorization

To uncover patterns in diagnostic data, several machine learning clustering techniques were applied, aiming to identify meaningful groupings of diagnoses that could simplify analysis and reveal trends in developmental conditions. The purpose of using these unsupervised clustering methods was to automatically group the same or similar diagnoses based on shared symptom characteristics, with the broader goal of enhancing the interpretability and structure of complex medical data.

A. K-Means Clustering

Clustering is a key data science technique for identifying patterns by grouping similar data points. It includes hierarchical and partitional methods, with k-means being a widely used

partitional approach. K-means partitions data into predefined clusters by minimizing intra-cluster variance but is sensitive to initialization and requires specifying the number of clusters in advance [9].

In this study, k-means was applied to group diagnoses based on symptom similarities. While the exact number of features used was not fixed, the clustering relied on extracted characteristics of symptoms, as these are the most relevant for capturing diagnostic similarity. However, due to the high variability and overlap of symptoms across conditions, the method failed to produce distinct, interpretable clusters. Many medical diagnoses exhibit broad, overlapping symptom spectrums, making it difficult for k-means to form well-separated groups. These challenges underscore its limitations for medical diagnosis classification and suggest the need for alternative clustering methods. The method's failure was evident in its inability to form clusters that aligned with known diagnostic categories, revealing that symptom-based similarity alone was insufficient for cleanly separating medical conditions.

B. Hierarchical Clustering

To address the limitations of k-means, hierarchical clustering with Ward's linkage method was tested. This approach builds a dendrogram, preserving hierarchical relationships between data points without requiring a predefined number of clusters. While this method provides a more flexible and structured clustering approach, the resulting clusters were still poorly defined, likely due to the complexity and overlapping nature of the symptoms. Again, failure was assessed by the interpretability and coherence of the resulting clusters. In this case, hierarchical clustering did not yield groupings that could be meaningfully mapped to clinical diagnoses, reinforcing the difficulty of categorizing developmental conditions using traditional clustering techniques alone.

C. DBSCAN Clustering

DBSCAN, a density-based clustering method, was explored as an alternative to k-means and hierarchical clustering. DBSCAN is particularly effective at identifying outliers and does not require a predefined number of clusters. It successfully detected rare conditions and isolated infrequent diagnoses. However, its sensitivity to density variations resulted in excessive noise labeling and difficulty forming coherent clusters for common conditions. This suggests that while DBSCAN is useful for identifying outliers, it is less effective for categorizing frequently occurring developmental disorders due to the diffuse and overlapping nature of symptom presentation.

D. Regex-Based Manual Mapping

Due to the limitations of unsupervised clustering, a regex-based manual classification approach was employed. This method, informed by medical literature and expert knowledge, mapped diagnoses into predefined categories based on keyword matching.

Regular expressions were used to identify specific diagnoses by detecting relevant terms in both English and local languages (e.g., "govor," "disfazija," "kašnjenje"). For example, autism spectrum disorders were identified using keywords such as "autism," "asperger," and "asd," while epilepsy-related diagnoses were matched with terms like "epilepsy," "infantile spasms," and "west syndrome." Diagnoses that did not match any predefined pattern were classified as "Other or Unclassified."

The regex-based approach successfully grouped diagnoses into meaningful categories, including Autism Spectrum (F84), Speech and Language Disorders (F80), and Down Syndrome (Q90.9). Compared to clustering methods, this approach was the most reliable, producing clear and interpretable classification results aligned with medical expertise.

VI. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to examine the dataset's structure, identify key trends, and assess data quality. Summary statistics were computed for numerical variables, while frequency distributions were analyzed for categorical variables.

The dataset includes participants from three countries, with Croatia having the highest representation. Household structures and child demographics were examined, revealing variations in family composition and the number of children with disabilities per household. A heatmap of missing values was generated to identify response patterns, guiding data imputation strategies.

The prevalence of developmental disabilities was analyzed, with conditions such as autism spectrum disorder (ASD), speech and language disorders, epilepsy, cerebral palsy, ADHD, and Down syndrome being the most frequently reported. The impact of these disabilities on families was assessed by examining service utilization, parental stress, and reported behavioral challenges. This analysis highlighted discrepancies in access to healthcare and support services, as well as variations in family needs and priorities.

A. Data Visualization

To better illustrate key findings and contextualize the dataset, a series of visualizations were created. These charts were specifically selected to represent critical aspects of the lives of children with developmental disabilities and their families, focusing on prevalence, household dynamics, support systems, and parental priorities.

Fig. 1 presents a bar chart showing the total number of children with disabilities per country. This graph was selected to emphasize the scale and distribution of disability across different national contexts. It helps identify regional disparities in diagnostic prevalence and reporting, with Croatia showing the highest number of diagnosed children.

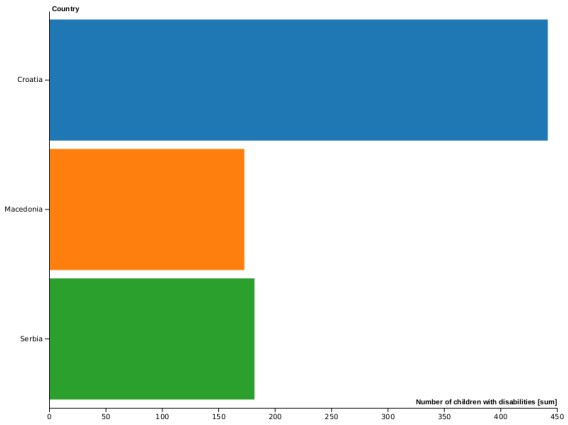


Fig. 1. Total Number of Children with Disabilities

Fig. 2 visualizes the number of children with disabilities per household. This chart was chosen to explore the family structures within the dataset and to provide insights into caregiving dynamics. The data reveals that most families have one child with a disability, but a significant number have two or more, indicating the varying degrees of caregiving burden.

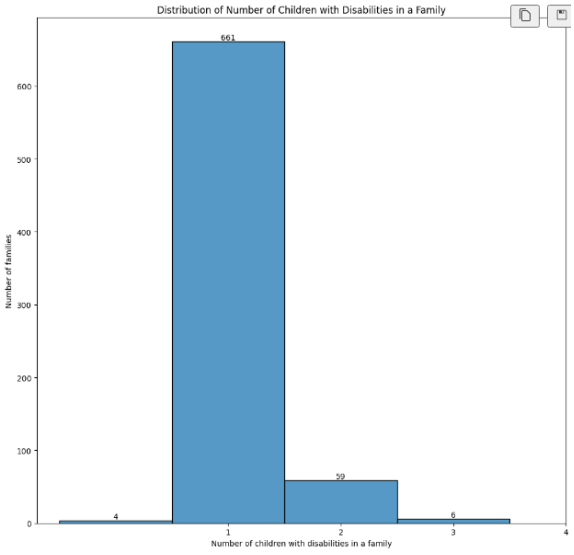


Fig. 2. Number of Children with Disabilities in the Families

Fig. 3 displays the priorities of families, including healthcare, education, psychological support, and disability rights, offering insights into their most pressing concerns. This visualization was selected to amplify the voices of parents and caregivers, offering a qualitative glimpse into their most urgent concerns. The emphasis on rights and health services suggests a broader need for systemic change, beyond just medical interventions, reflecting the multidimensional nature of disability-related challenges.

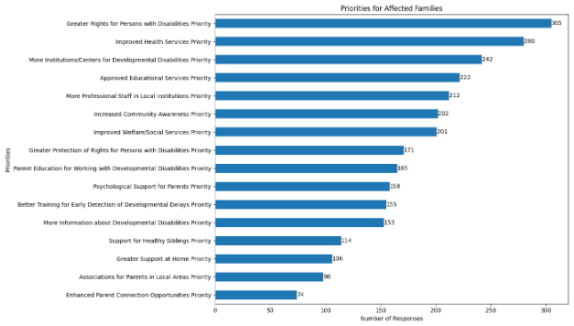


Fig. 3. Priorities for Affected Families

Fig. 4 illustrates the frequency of different types of assistance received by families. This graph was chosen to assess the landscape of support systems, from government aid to local and community-based services. It reveals gaps in service distribution, with national-level government assistance being the most common and religious or municipal support being nearly absent. These findings help highlight opportunities for more equitable and diversified support channels.

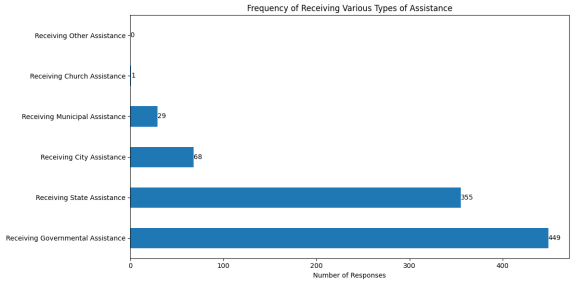


Fig. 4. Frequency of Receiving Various Types of Assistance

These visualizations offer valuable insights into key aspects of the dataset, including the prevalence of disabilities, family structures, and the types of support families receive. They highlight areas where services could be improved, particularly in terms of accessibility, availability, and effectiveness. By examining these visual insights, we gain a better understanding of the challenges faced by caregivers, emphasizing the need for improved resource allocation, targeted interventions, and policy adjustments to better support families of children with disabilities.

VII. Conclusion

This study presents a detailed account of the preprocessing and exploratory analysis of a dataset on developmental disabilities collected through the Erasmus+ SynergyEd project. The dataset underwent extensive cleaning, standardization, and normalization to ensure consistency and accuracy. Machine learning techniques were tested for diagnosis categorization, but manual regex-based classification proved to be the most effective method. Exploratory data analysis provided valuable

insights into country distributions, diagnostic trends, and service utilization patterns.

This work lays the groundwork for future predictive modeling and a deeper understanding of the challenges faced by children with developmental disabilities. Future research should focus on developing predictive models to enhance diagnostic classification and improve service provision for these children.

REFERENCES

[1] J. Meaković, A. Čivljak, T. Zorčec, V. Vučić, D. Ristić-Medić, A. Veselinović, M. Čivljak, and L. Puljak, "Perceptions on support, challenges and needs among parents and caregivers of children with developmental disabilities in croatia, north macedonia and serbia: a cross-sectional study," *BMC pediatrics*, vol. 24, no. 1, p. 297, 2024.

[2] Erasmus+ SynergyEd Project, "Perceptions of support, challenges, and needs among parents/caregivers of children with developmental disabilities," 2024. [Online]. Available: https://osf.io/a9w8z

[3] S. Shukla, "Unlocking the power of data: An introduction to data analysis in healthcare," *Int J Comput Sci Eng*, vol. 11, no. 3, pp. 1–9, 2023.

[4] A. D. Gupta, K. Singh, K. D. Singh, P. Kushwaha, B. P. Lohani, and S. Kumar, "Unveiling insights: Exploring healthcare data through data analysis," in *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*. IEEE, 2024, pp. 575–581.

[5] A. A. Lydia, M. Poongundran, S. Masarath, P. V. Karthick, I. M. Zayats *et al.*, "Eda and its impact in dataset discover patterns in the service sector," in *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2022, pp. 940–945.

[6] S. Almuhaideb and M. E. B. Menai, "Impact of preprocessing on medical data classification," *Frontiers of Computer Science*, vol. 10, pp. 1082–1102, 2016.

[7] B. M. Konopka, F. Lwow, M. Owczarz, and Ł. Łaczmański, "Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data," *PloS one*, vol. 13, no. 8, p. e0201950, 2018.

[8] H. W. Dhany, S. Sutarman, and F. Izhari, "Exploratory data analysis (eda) methods for healthcare classification," *Journal of Intelligent Decision Support System (IDSS)*, vol. 6, no. 4, pp. 209–215, 2023.

[9] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE access*, vol. 8, pp. 80 716–80 727, 2020.

AI Based Cyberbullying Detection - State of the Art

Mahnoor Jamil*
Faculty of Computer Science and Engineering
mahnoor.jamil@students.finki.ukim.mk
Ss. Cyril and Methodius University
Skopje 1000, North Macedonia

Ivan Chorbev
Faculty of Computer Science and Engineering
ivan.chorbev@finki.ukim.mk
Ss. Cyril and Methodius University
Skopje 1000, North Macedonia

Boban Joksimoski
Faculty of Computer Science and Engineering
Boban.joksimoski@finki.ukim.mk
Ss. Cyril and Methodius University
Skopje 1000, North Macedonia

***Abstract*—Cyberbullying on social media has become a widespread problem, that resulted in emotional distress and mental health problems for social media users. Artificial Intelligence (AI) offered a promising solution to this problem by enabling the identification of abusive language patterns in huge amounts of content. This research is focused on providing comparative analysis to highlight which AI model has achieved highest accuracy. Support Vector Machine (SVM), Logistic Regression, Random Forest, Naive Bayes, and K-Nearest Neighbour (KNN) are some of the approaches that have been studied for the detection of cyberbullying; each has its own benefits and limitations. The goal of this study is to determine the accuracy and robustness of these algorithms in detecting cyberbullying on platforms like Wikipedia and Twitter. The findings will contribute to the utilisation of dependable and effective algorithm for preventing cyberbullying.**

***Index Terms*—Cyberbullying, Artificial Intelligence, Abusive Content Detection, Comparative Analysis**

I. Literature Review

A. Research Background

Previous studies have explored using AI for detecting abusive content in social media. [3] expanded the scope by incorporating both Bangla and Romanized Bangla datasets, totaling 12,000 samples, and employed Multinomial Naive Bayes and Convolutional Neural Networks (CNN). This study underscores the importance of language-specific models and the potential of deep learning techniques in multilingual contexts. In a more recent study, [4] explored multiple platforms, including social media, forums, and websites, using a dataset of 10,266 samples in Bangla and Chittagonian. They employed a diverse range of techniques, from traditional machine learning (e.g., SVM, Random Forest) to deep learning models (e.g., CNN, LSTM, GRU) and hybrid approaches, showcasing the evolution of cyberbullying detection methods.

As highlighted in Table I, [5] analyzed Twitter data with 37,373 tweets, likely in English, and tested algorithms such as Logistic Regression, Random Forest, and SVM. Their findings align with other studies in highlighting the effectiveness of ensemble methods and traditional algorithms for structured social media data. [6] also focused on Twitter, using 28,357 samples and employing algorithms like Naive Bayes, Random Forest, and SVM. Their work reinforces the importance of algorithm selection based on dataset characteristics.

[7] shifted focus to Wikipedia, analyzing 275,553 English comments and utilizing a mix of traditional machine learning (e.g., XGBoost, SVM) and deep learning models (e.g., LSTM, Bi-GRU). This study highlights the adaptability of detection methods to different platforms and content types.

Finally, [8] combined Twitter and Formspring data, using 37,373 tweets and testing algorithms such as Logistic Regression, LightGBM, and SVM. Their work underscores the potential of cross-platform data integration to enhance detection accuracy. Collectively, these studies illustrate the diverse methodologies and challenges in cyberbullying detection across platforms and languages.

Study	Platform(s)	Dataset Size	Algorithm Used
[3]	Bangla, Romanized Bangla	12,000 samples	Multinomial Naive Bayes, CNN
[4]	Social Media, Forums, Websites	10,266 samples	SVM, Random Forest, CNN, LSTM, GRU
[5]	Twitter	37,373 tweets	Logistic Regression, Random Forest, SVM
[6]	Twitter	28,357 samples	Naive Bayes, Random Forest, SVM
[7]	Wikipedia	275,553 comments	XGBoost, SVM, LSTM, Bi-GRU
[8]	Twitter, Formspring	37,373 tweets	Logistic Regression, LightGBM, SVM

TABLE I
OVERVIEW OF PREVIOUS STUDIES ON CYBERBULLYING DETECTION

B. Search Process

The research involved searching through multiple scientific databases, such as Google Scholar, Semantic Scholar, IEEE, ResearchGate, MDPI, and Scopus. A total of 50 papers were reviewed, with 38 excluded for being off-topic. The remaining 12 papers were thoroughly analyzed, particularly those that explored the intersection of AI and social media content moderation.

C. Searching Criteria Definition

The papers were selected based on specific evaluation criteria, including the relevance to topics like:

- **Cyberbullying and Harassment in Social Media**
- **AI Applications in Social Media Monitoring**

The review primarily focused on AI-based models for detecting harmful content on social media platforms.

II. MODEL TRAINING PROCESS

Figure 1 highlights the process of training a model to give accurate results [12]. Training a model initiates with collecting relevant data for the task and preprocessing it to clean and prepare it, which may include handling missing values and scaling features.

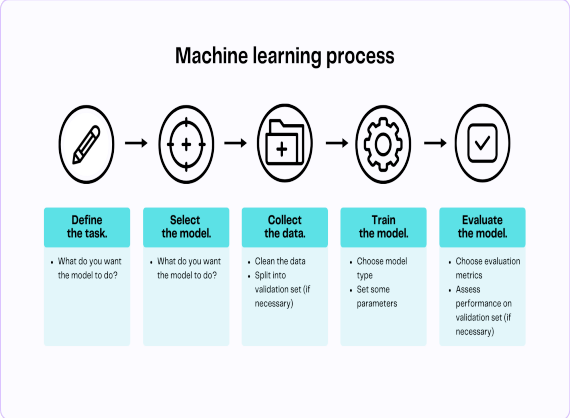


Fig. 1. Model Training Process

Furthermore, the data is split into training and testing sets,. Later, appropriate model are trained using the training data to allow the model to learn patterns. After training, evaluation of the model's performance is conducted using the testing set to assess its accuracy and effectiveness.

III. AI APPROACHES FOR ABUSIVE CONTENT DETECTION

AI has increasingly been applied to social media content analysis, especially for identification of abusive language. Multiple AI methodologies have been used to develop systems that can detect offensive behavior. The following models have shown great potential in the detection of harmful content on Twitter:

1) Random Forest: This algorithm is effective at handling high-dimensional datasets, making it suitable for classification tasks such as toxicity detection. Random forest reduces overfitting by averaging the results of multiple decision trees, thereby improving the model's ability to generalise. In the context of cyberbullying detection, Random Forest leverages its ability to capture labels of abusive content, providing better performance and robustness to outliers [9].

2) Logistic Regression: In order to map the link between input features and the likelihood of a specific class label, logistic regression applies a function. Using vectorized textual properties, Logistic Regression

can effectively categorize text samples as abusive or non-abusive. However, when working with intricate language patterns, performance might be constrained [10].

3) K-Nearest Neighbor: A non-parametric, instance-based learning technique called the KNN algorithm groups data points according to the majority class of their k nearest neighbors in the feature space. KNN is a popular option for preliminary testing in text classification problems since it is simple and easy to understand. When it comes to detecting cyberbullying, KNN classifies a tweet or text sample according to how closely its attributes resemble those of labeled samples in the training dataset. However, because of the curse of dimensionality and the increasing computing complexity, its efficacy may decrease with high-dimensional data, like textual inputs [11].

4) Support Vector Machines: SVMs are frequently employed for categorization problems. SVMs identify the ideal hyperplane to distinguish abusive from non-abusive content in order to detect abusive content [1]. When there is a distinct difference between classes, like when separating hate speech from neutral language, SVMs perform especially well. SVMs provide excellent performance in toxicity detection when paired with word representations such as TF-IDF (Term Frequency-Inverse Document Frequency).

5) Naive Bayes Classifier: A popular baseline model for text classification is the Naive Bayes classifier, based on probabilistic reasoning. Naive Bayes has gained success in abusive content detection tasks despite its simplicity, especially when the language is clear and the data contains little noise [2]. Naive Bayes is simple to use and understand because it determines the likelihood that each word will appear in abusive and non-abusive circumstances.

IV. Data Processing

TweetEval is a comprehensive dataset designed for NLP research on Twitter data by Huggingface. It includes a variety of tasks, such as sentiment analysis, emoji prediction, and toxin detection. The toxicity detection subset, which included tagged tweets classified as abusive or non-abusive, was used for this investigation. Robust model training is made possible by the dataset's balanced combination of benign tweets and real-world abusive content. Its multilingual support also guarantees flexibility in a variety of linguistic environments, such as Arabic, Spanish, and English. This multilingual feature makes it possible to train the model on a variety of data and improves its ability to identify harmful behavior in a broad range of linguistic and cultural contexts. Additionally, TweetEval is regularly updated to guarantee that the dataset is current and reflective of social media discourse.

V. Measuring Performance Metrics

The AI model's performance is assessed using standard measures including F1-score, recall, accuracy, and precision. These measurements shed light on how well the algorithm detects abusive content and reduces false positives and negatives. A labeled dataset of known abusive and non-abusive tweets is used to validate the AI model's output.

VI. Results

Several machine learning algorithms' effectiveness in detecting cyberbullying has been assessed on a variety of platforms, most notably Wikipedia and Twitter. The Table II shows variation in accuracy with respect to training of different models. Depending on the dataset and feature set, SVM algorithms have shown an average accuracy of 92%. On Twitter datasets, Random Forest algorithms have shown greater accuracy, averaging about 93%. They can handle noisy and unstructured social media data, which is typical in cyberbullying situations, because to their ensemble-based design. Experiments on Twitter have shown that models of Logistic Regression have an average accuracy of 90%. Naive Bayes classifiers have shown a comparatively lower average accuracy of 83%. However, KNN classifiers performed poorly in comparison, with an average accuracy of only 27%, indicating limited applicability in text data environments like Twitter.

VII. Discussion

In this study, we analyzed effectiveness of multiple machine learning algorithms from several cyberbullying detection experiments. Four algorithms— SVM, Random Forest, and Logistic Regression—repeatedly attained high accuracy levels between 90% and 92%, according to our data. With a range of more than 20 percentage points, SVM and Random Forest had the greatest variability in the performance range analysis, indicating that feature engineering and the dataset have a significant impact

Machine Learning Algorithm	Platform	Accuracy
Support Vector Machine	Wikipedia/Twitter	92%
Random Forest	Twitter	93%
Logistic Regression	Twitter	90%
Naive Bayes	Twitter	83%
K-Nearest Neighbors	Twitter	27%

TABLE II
ACCURACY RESULTS OF MACHINE LEARNING ALGORITHMS IN DETECTING CYBERBULLYING

on how effective it is. With a range of fewer than 10 percentage points, Naive Bayes, on the other hand, showed more consistent results. The investigation also shows that simpler models like KNN perform poorly because of its constraints in handling complex text data.

VIII. CONCLUSION

In conclusion, our examination of several machine learning methods for the identification of cyberbullying shows that SVM and Logistic Regression, consistently exhibit excellent accuracy and stability. The most dependable algorithm for this work is Random Forest, which successfully manages the complexity of social media text data with the highest average accuracy of 93.63%.

The need of employing more robust models for such complex tasks is highlighted by the noticeably low performance of simpler models like KNN. To improve these algorithms' practicality, future studies should investigate more optimization and expand testing to other platforms.

REFERENCES

[1] Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.
[2] McCallum A., Nigam K. "A Comparison of Event Models for Naive Bayes Text Classification." Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, 1998.
[3] Ahmed M. T., Rahman M., Nur S., Islam A., Das D. "Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla Text: A Comparative Study." 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2021.
[4] Mahmud T., Ptaszynski M., Masui F. "Exhaustive Study into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts." Electronics, 2024.
[5] Muneer A., Fati S. "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter." Future Internet, 2020.
[6] Nethra D., Suthisini P., Merlien J. J., Shanmugapriya D., Padmavathi G., Roshni A. "Performance Analysis of Cyberbullying Threats in Social Media Using Machine Learning Techniques." International Conference on Computing for Sustainable Global Development, 2024.
[7] Raj C., Agarwal A., Bharathy G., Narayan B., Prasad M. "Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques." Electronics, 2021.
[8] Sahana V., A. K. M., Darem A. A. "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on FormSpring in Textual Modality." International Journal of Computer Network and Information Security, 2023.
[9] Breiman L. "Random Forests." Machine Learning, 45(1), 5–32, 2001.
[10] Hosmer D. W., Lemeshow S., Sturdivant R. X. "Applied Logistic Regression." John Wiley and Sons, 2013.
[11] Cover T., Hart P. "Nearest Neighbor Pattern Classification." IEEE Transactions on Information Theory, 13(1), 21–27, 1967.
[12] Newhauser, M. "Five Diagrams to Understand AI." GP Tech Blog, 17 Oct 2023. Available at: https://www.gptechblog.com/5-diagrams-to-help-you-understand-generative-ai/

Large Language Models for Legal Reasoning in Low Resource Languages

Ivan Apostolovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, North Macedonia
ivan2002mk@gmail.com

Martin Angelovski
Amazon UK
London, United Kingdom
mangelovski@gmail.com

Vladimir Radevski
School of Computer Science and Information Technology
University American College Skopje
Skopje, North Macedonia
vladimir.radevski@uacs.edu.mk

Marija Stankova Medarovska
School of Computer Science and Information Technology
University American College Skopje
Skopje, North Macedonia
marija.s.medarovska@uacs.edu.mk

***Abstract*—Legal Artificial Intelligence (LegalAI) has advanced significantly with the development of Large Language Models (LLMs). However, these models face challenges in legal reasoning due to limited domain knowledge and a tendency to generate fabricated responses. This issue is particularly pronounced in low-resource languages, such as Macedonian language, where digitized legal corpora are scarce. Most prior studies have focused on high resource languages, despite legal practices varying across jurisdictions, each with distinct sources of law and legal traditions. This disparity limits the potential of LLMs to support legal professionals and improve access to justice in underrepresented legal systems. The study explores the adaptation of LLMs for Macedonian legal text processing using Retrieval-Augmented Generation (RAG). We construct a novel dataset sourced from the Official Gazette of North Macedonia and implement RAG enhanced LLMs to assess their legal text comprehension and question-answering performance. The study compares different embedding models (Amazon Titan, Cohere) and LLM architectures (Llama 3.2, Claude 3 Haiku, Amazon Nova Pro) using automated evaluation metrics. The findings suggest that embedding selection and model architecture significantly impact response quality, with Cohere embeddings and Amazon Nova models demonstrating better semantic alignment and retrieval accuracy. The study contributes to the advancement of LegalAI by providing a scalable framework for adapting LLMs to low resource languages and improving access to legal technology in underrepresented jurisdictions.**

Keywords—Artificial Intelligence, Large Language Models, Retrieval Augmented Generation, Legal AI

I. INTRODUCTION

The field of Legal Artificial Intelligence (LegalAI) focuses on applying artificial intelligence (AI) techniques to the legal domain [1]. As most of the resources in the domain are primarily text based, such as court rulings, agreements and legislation, LegalAI particularly relies on natural language processing (NLP) techniques. In the past several years, Large Language Models (LLMs), built on transformer architectures, have transformed NLP by achieving remarkable performance across various linguistic tasks [2, 3]. These models are trained on large corpora of text data, mainly sourced from the web, and have the objective of predicting the next word in a partial sentence [3]. General LLMs, however, face difficulties with legal language and reasoning due to their limited domain knowledge and comprehension of legal syntax and vocabulary [4, 5, 8]. Moreover, they may generate fabricated responses, commonly known as hallucinations [6]. These limitations come from the LLMs reliance on training data, which can cause issues when the queries require information beyond the training base or up to date knowledge [7]. Currently, legal analysis is performed mainly by experts with years of specialized training. Enhancing LLMs with legal capabilities could not only improve the efficiency of legal professionals, but also help meet the growing demand for legal assistance, ultimately improving public access to justice [8].

The most commonly used methods for improving the performance of LLMs on domain specific topics are fine-tuning (FT) and Retrieval-Augmented Generation (RAG) [7, 9]. FT involves updating the parameters of a pre-trained LLM, by continuing the training on a smaller, specialized corpus [10]. While this approach enhances the model's domain knowledge, it requires significant computational resources and frequent retraining to incorporate new information. RAG, on the other hand, is a more cost efficient method for incorporating new data into the LLM, without the need to retrain the model [11]. Instead of relying only on pre-trained knowledge, RAG dynamically retrieves relevant information from an external knowledge base before generating responses, which leads to more accurate outputs. By directly citing the sources of information, RAG allows legal experts to verify the retrieved information and generated responses before using it in legal reasoning.

While initial progress has been made in adapting general LLMs to the legal domain, the research has been mainly dedicated to high-resource languages, such as English [12, 13] and Chinese [8, 14, 15, 31]. However, legal practice varies across jurisdictions, as they have distinct sources of law and legal traditions, such as common law and civil law systems 16]. As a result, low-resource languages, such as Macedonian language, remain largely underexplored due to limited digital resources and the linguistic diversity in available datasets [17, 18]. This disparity restricts the potential of LLMs to support legal professionals and improve access to justice in underrepresented legal systems [19].

This study addresses the gap, by adapting LLMs for Macedonian legal text processing and leveraging Retrieval-Augmented Generation (RAG) to enhance accuracy and

reduce fabricated responses. By constructing a novel dataset sourced from the Official Gazette of North Macedonia, we evaluate the effectiveness of RAG enhanced LLMs in legal question-answering and text comprehension tasks. Through a comparative analysis of embedding models (Amazon Titan, Cohere) and LLM architectures (Llama 3.2, Claude 3 Haiku, Amazon Nova Pro), we assess the impact of model and embedding selection on response quality. The findings indicate that RAG can significantly improve LLM performance in the legal domain within a low-resource language setting. By proposing a replicable framework for integrating retrieval-based techniques with LLMs, this study aims to provide a basis for future research on legal reasoning in underrepresented languages and jurisdictions.

The remainder of this study is structured as follows. Section 2 discusses the dataset and methodology, including the implementation of RAG and evaluation metrics. Section 3 presents the experimental results, comparing different embeddings and LLM architectures. Section 4 provides a discussion of the findings, highlighting the implications and limitations of the study. Section 5 provides concluding remarks.

II. Materials and Methods

A. Data

The dataset used in this study is sourced from the Official Gazette of the Republic of North Macedonia, which is the official national publication for legal and administrative information. All issues are publicly available on the PE Official Gazette of the Republic of North Macedonia website [20]. The data constitute a comprehensive corpus of Macedonian language legal texts, including laws, decrees and public notices spanning multiple years. For the study, only the issues from the year 2024 were considered, which is a total of 278 articles.

B. Retrieval-Augmented Generation (RAG)

LLMs have demonstrated remarkable capabilities on many NLP tasks, however, they face significant challenges such as tendency to generate fabricated responses, difficulty of incorporating new information and a lack of transparency in responses [7]. In order to address these issues and improve domain specific performance, commonly used techniques are fine-tuning (FT) and Retrieval-Augmented Generation (RAG) [7, 9]. LLMs store their knowledge obtained from the training data in the parameters of the model, also known as parametric knowledge. The first approach, FT, involves adapting the parameters of a pre-trained LLM, by continuing its training on a smaller and specialized corpus [10]. While effective, this approach is computationally intensive and costly, making it impractical for domains that require frequent updates, such as the legal field. For instance, the Official Gazette of the Republic of North Macedonia publishes multiple issues per month, which requires continuous updates to ensure the LLM is based on accurate and up to date legal knowledge. The second approach, RAG, offers a more flexible alternative by leveraging non-parametric knowledge stored in external sources such as vector databases, which can be updated dynamically without modifying the model itself [11]. Research suggests that integrating non-parametric knowledge into LLMs allows smaller models to perform at levels comparable to larger ones [11]. Furthermore, the study of Soudani et al. [9] indicates that RAG significantly outperforms FT when applied to the same LLM, particularly when dealing with less commonly available factual knowledge. This is relevant for low resource languages like Macedonian, where legal knowledge has less digital resources.

Fig. 1 presents the implementation process of RAG, which consists of several stages. Initially, the legal documents from the Official Gazette are converted into embeddings, which are numerical representations that capture the semantic meaning of text and stored in a vector database. In this study, we experiment with two embedding methods: Amazon Titan Text Embedding (amazon.titan-embed-text-v2:0) [21] and Cohere Multilingual Embedding Model v2.0.1 [22]. According to their official documentation, both models support Macedonian language.

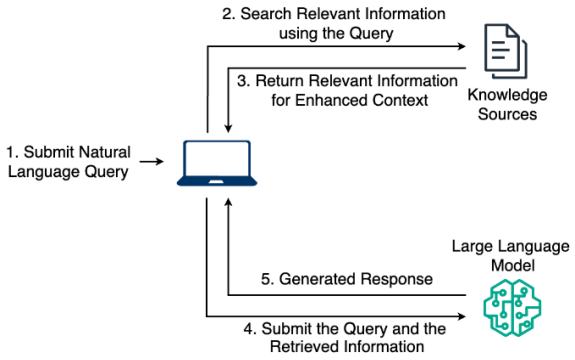


Fig. 1. Retrieval-Augmented Generation (RAG) workflow.

Once the database is created, a user can submit a natural language query, which prompts an orchestrator to conduct a similarity search within the vector database to retrieve the most relevant information. The retrieved data is then incorporated into the query prompt before being processed by the LLM, which generates a response by leveraging both parametric knowledge and the retrieved external context (non-parametric knowledge). For LLMs, this study examines three models: Llama 3.2 11B which is a multimodal large language models with 11 billion parameters by Meta [23], Claude 3 Haiku by Anthropic [24] and Amazon Nova Pro by Amazon [25] with an undisclosed number of parameters. In the legal domain it is important that LLM generated responses are accurate, reliable and aligned with the relevant legal frameworks. RAG enhances these aspects by retrieving authoritative legal information, which reduces the risk of hallucinations. By examining the retrieved data, legal professionals can verify the quality of the generated responses.

C. Evaluation

In order to evaluate the RAG framework in the context of the Macedonian legal domain, we conduct an end-to-end RAG evaluation [7]. This evaluation assesses the final response generated by the LLM, which integrates non-parametric knowledge from an external database, by comparing it to the input query and the ground truth, i.e., the correct answer to the query. Assessing the quality of the responses through human annotation can be costly and impractical for evaluating an entire corpus [26]. Furthermore, test data in the Macedonian legal domain is limited. Therefore, we employ automated evaluation metrics on a carefully constructed test set from the Official Gazette that consists of 100 factual question-answer examples. Factual questions and answers are suited for testing

RAG, as they can be directly validated against the Official Gazette articles, without needing external context.

The first set of evaluation measures is based on the RAG evaluation framework RAGAS [27]. RAGAS relies on strong LLMs, such as GPT-4o in our case, to act as impartial evaluators ("judges") of other LLM generated responses. This approach aligns with recent research suggesting that LLMs can serve as effective evaluators for subjective and qualitative tasks. For example, the study of Zheng et al. [28] indicates that LLM based evaluators perform comparably to human evaluators in controlled assessments. More specifically, we consider the RAGAS evaluation metrics Faithfulness, which measures the factual consistency of the generated response and Answer Relevance, which is the accuracy of the generated response when compared to the ground truth.

Furthermore, we also use two additional evaluation metrics to assess the quality of the generated responses by measuring their similarity to the ground truth. LaBSE (Language-Agnostic BERT Sentence Embedding), combined with cosine similarity, measures semantic similarity at the sentence level by computing the cosine distance between their embeddings [29]. The model is developed by Google and supports more than 100 languages including Macedonian language. On the other hand, BERTScore F1 measure evaluates text similarity using a token based approach by comparing contextual word embeddings [30]. The two metrics complement each other by capturing different aspects of text similarity. LaBSE measures sentence level semantic similarity that is useful for paraphrased content, such as one generated by probabilistic LLMs. Meanwhile, BERTScore F1 captures token level contextual similarity, detecting fine grained differences in phrasing.

III. Results

The evaluation of different Large Language Models (LLMs) and embedding techniques reveals notable differences in model performance across the evaluation metrics: LaBSE, BERTScore F1 and the faithfulness and answer relevancy metrics from the RAGAS framework.

Table 1 presents a detailed comparison of the evaluation results among the different models. The combination of using Cohere embeddings with Amazon Nova LLM achieved the highest scores in LaBSE similarity (0.77) and BERTScore F1 (0.82), indicating that its generated responses have good semantic alignment with the reference answers, at both the sentence and token levels. Cohere Nova also achieved the highest results in faithfulness (0.42) and answer relevancy (0.71), suggesting that it produced the most contextually appropriate and factually grounded responses. In contrast, Titan Llama had the highest variability across most metrics, which indicates inability to generalize well across different types of queries. The scores are overall lower for the RAGAS metrics, which should be interpreted with caution, as they also reflect the understanding and interpretation of the legal context in Macedonian language of the evaluator LLM (GPT-4o). Although GPT-4o is a strong model [3], its evaluation may be affected by limited exposure to Macedonian legal texts in its training data. On the other hand, LaBSE and BERTScore F1 are measures based on text similarity and semantic relevance, and their higher scores indicate that the retrieved documents and generated answers are semantically close to the query and context, which is important for legal accuracy.

TABLE I. Comparison of the evaluation metric scores among the different combinations of embeddings and LLMs.

LLM + Embedding	LaBSE		BERTScore		Faitfulness		Answer Rel.	
	Mean	***SD***	***Mean***	***SD***	***Mean***	***SD***	***Mean***	***SD***
Cohere Claude	0.74	0.13	0.80	0.08	0.35	0.35	0.69	0.17
Cohere Llama	0.76	0.18	0.80	**0.11**	0.41	0.39	0.67	0.21
Cohere Nova	**0.77**	0.13	**0.82**	0.08	**0.42**	0.41	**0.71**	0.13
Titan Claude	0.71	0.15	0.78	0.08	0.32	0.36	0.66	0.22
Titan Llama	0.73	**0.21**	0.80	**0.11**	0.39	**0.42**	0.59	**0.29**
Titan Nova	0.75	0.15	0.80	0.09	0.42	0.42	0.65	0.23
Average	**0.74**	**0.16**	**0.80**	**0.09**	**0.39**	**0.40**	**0.66**	**0.21**

To further analyze the impact of individual LLMs and embedding models, the data was aggregated and presented in Tables 2 and 3, respectively. Table 2 aggregates the performance based on the embeddings used, averaging the scores of Cohere and Titan models across different LLMs. From the results, Cohere embeddings outperform Titan embeddings across all four evaluation metrics, with Cohere models achieving the highest overall faithfulness (0.39) and answer relevancy (0.69). The LaBSE similarity score (0.76) and BERTScore F1 (0.81) for Cohere were also higher than Titan, indicating that Cohere's responses were more semantically aligned with the references. Titan models, on the other hand, exhibited higher variability, as reflected in the standard deviation values.

Table 3 compares the three LLMs, Claude, Llama and Nova, independently of embeddings. Amazon Nova achieved the highest LaBSE similarity (0.76), BERTScore F1 (0.81) and faithfulness (0.42), outperforming both Claude and Llama. On the answer relevancy metric, it achieved the same result as Claude (0.68). This consistent performance of Nova across both semantic and factual dimensions suggests that the LLM is more effective at integrating retrieved Macedonian legal content into its outputs than the other two LLMs. Llama, on the other hand, had the highest variability in three out of the four categories. This variability suggests that while Llama is capable of producing high quality responses in some cases, its outputs are less predictable and may fluctuate significantly depending on the input query. In the context of legal applications, such inconsistency poses a challenge, as legal reasoning requires a high degree of reliability.

TABLE II. Aggregated evaluation metric scores on an embedding level

Embedding	LaBSE		BERTScore		Faitfulness		Answer Rel.	
	Mean	***SD***	***Mean***	***SD***	***Mean***	***SD***	***Mean***	***SD***
Cohere	**0.76**	0.15	**0.81**	**0.09**	**0.39**	0.38	**0.69**	0.17
Titan	0.73	**0.17**	0.79	**0.09**	0.38	**0.40**	0.63	**0.25**
Average	**0.74**	**0.16**	**0.80**	**0.09**	**0.38**	**0.39**	**0.66**	**0.21**

TABLE III. Aggregated evaluation metric scores on an LLM level.

LLM	LaBSE		BERTScore		Faitfulness		Answer Rel.	
	Mean	***SD***	***Mean***	***SD***	***Mean***	***SD***	***Mean***	***SD***
Claude	0.73	0.14	0.79	0.08	0.34	0.36	**0.68**	0.20
Llama	0.75	**0.20**	0.80	**0.11**	0.40	0.41	0.63	**0.25**
Nova	**0.76**	0.14	**0.81**	0.09	**0.42**	**0.42**	**0.68**	0.18
Average	**0.75**	**0.16**	**0.80**	**0.09**	**0.39**	**0.40**	**0.66**	**0.21**

These results suggest that model choice and embedding selection significantly influence response quality, with Cohere embedding and Amazon Nova among the LLMs generating the most consistent and accurate results.

IV. Discussion

This study contributes to the advancement of Legal AI by adapting LLMs for legal text comprehension in Macedonian language, which is an underrepresented low resource language. The use of RAG, combined with different embedding and LLM models, presents a scalable approach that may benefit other low resource legal domains with similar linguistic and data constraints. The study, however, has several limitations. First, the reliance on automated evaluators, such as GPT-4o, in the RAGAS framework introduces the risk of subjectivity. The evaluation of faithfulness and answer relevance could be improved by involving legal professionals to verify LLM generated responses for legal accuracy, even on a smaller sample. Second, the dataset is based on the Official Gazette of North Macedonia and could be expanded to include additional sources, such as court rulings and contracts, to provide a more comprehensive representation of Macedonian legal reasoning. Furthermore, the testing phase of this study focused on factual retrieval, using factual question-answer examples from the dataset. Future research could introduce reasoning-based tasks, to assess deeper legal reasoning capabilities in the Macedonian legal context.

V. Conclusion

The paper explores the adaptation of Large Language Models (LLMs) for legal reasoning in Macedonian language, by leveraging Retrieval-Augmented Generation (RAG). Our findings suggest that embedding selection and model architecture play an important role in LLM performance for legal tasks. Cohere embeddings and Amazon Nova based models showed stronger semantic alignment and better retrieval accuracy than other configurations. However, the generally lower faithfulness scores across all models are concerning and emphasize the need for human oversight and expanded datasets to enhance reliability. Overall, the findings suggest that LLMs, when supported by retrieval-based methods, have the potential to assist legal professionals in low resource settings. With continued refinement, these techniques may help narrow the gap between legal AI development in high resource and low resource languages, contributing to more accessible and reliable legal information in underrepresented jurisdictions. Future work could incorporate human evaluation of the tasks by legal professionals, expand the dataset to include court rulings, contracts and additional legal sources and benchmark RAG models against fine-tuned LLMs.

References

[1] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z. and Sun, M., 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. arXiv preprint arXiv:2004.12158.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30, pp.5998-6008.

[3] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S. and Nori, H., 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712

[4] Chalkidis, I., Androutsopoulos, I. and Aletras, N., 2019. Neural legal judgment prediction in English. arXiv preprint arXiv:1906.02059.

[5] Kandpal, N., Deng, H., Roberts, A., Wallace, E. and Raffel, C., 2023, July. Large language models struggle to learn long-tail knowledge. In International Conference on Machine Learning (pp. 15696-15707). PMLR.

[6] Zhang, Y., Cui, L., Bi, W. and Shi, S., 2023. Alleviating hallucinations of large language models through induced hallucinations. arXiv preprint arXiv:2312.15710.

[7] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H. and Wang, H., 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2, p.1.

[8] Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z. and Ge, J., 2023. Lawbench: Benchmarking legal knowledge of large language models. arXiv preprint arXiv:2309.16289.

[9] Soudani, H., Kanoulas, E. and Hasibi, F., 2024, December. Fine tuning vs. retrieval augmented generation for less popular knowledge. In Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (pp. 12-22).

[10] Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L., 2023. Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 36, pp.10088-10115.

[11] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, pp.9459-9474.

[12] Colombo, P., Pires, T.P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A.F., Esposito, F., Raposo, V.L., Morgado, S. and Desa, M., 2024. Saullm-7b: A pioneering large language model for law. arXiv preprint arXiv:2403.03883.

[13] Shu, D., Zhao, H., Liu, X., Demeter, D., Du, M. and Zhang, Y., 2024, October. LawLLM: Law large language model for the US legal system. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 4882-4889).

[14] Yue, S., Liu, S., Zhou, Y., Shen, C., Wang, S., Xiao, Y., Li, B., Song, Y., Shen, X., Chen, W. and Huang, X., 2024, July. LawLLM: Intelligent Legal System with Legal Reasoning and Verifiable Retrieval. In International Conference on Database Systems for Advanced Applications (pp. 304-321). Singapore: Springer Nature Singapore.

[15] Xiao, C., Hu, X., Liu, Z., Tu, C. and Sun, M., 2021. Lawformer: A pre-trained language model for chinese legal long documents. AI Open, 2, pp.79-84.

[16] Henderson, P., Krass, M., Zheng, L., Guha, N., Manning, C.D., Jurafsky, D. and Ho, D., 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. Advances in Neural Information Processing Systems, 35, pp.29217-29234.

[17] Dobreva, J., Pavlov, T., Mishev, K., Simjanoska, M., Tudzarski, S., Trajanov, D. and Kocarev, L., 2022, September. MACEDONIZER-The Macedonian Transformer Language Model. In International Conference on ICT Innovations (pp. 51-62). Cham: Springer Nature Switzerland.

[18] Joshi, P., Barnes, C., Santy, S., Khanuja, S., Shah, S., Srinivasan, A., Bhattamishra, S., Sitaram, S., Choudhury, M. and Bali, K., 2019. Unsung challenges of building and deploying language technologies for low resource language communities. arXiv preprint arXiv:1912.03457.

[19] Savelka, J., Ashley, K.D., Gray, M.A., Westermann, H. and Xu, H., 2023. Explaining legal concepts with augmented large language models (gpt-4). arXiv preprint arXiv:2306.09525.

[20] PE Official Gazette of the Republic of North Macedonia, www.slvesnik.com.mk, 2025, Accessed:01/04/2025

[21] Amazon Titan Text Embeddings models, docs.aws.amazon.com/bedrock/latest/userguide/titan-embedding-models.html, 2025, Accessed:01/04/2025

[22] Unlocking the Power of Multimodal Embeddings, docs.cohere.com/v2/docs/multimodal-embeddings, 2025, Accessed:01/04/2025

[23] Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).

[24] Introducing the next generation of Claude, www.anthropic.com/news/claude-3-family, 2025, Accessed:01/04/2025

[25] Amazon Nova Foundation Models, aws.amazon.com/ai/generative-ai/nova, 2025, Accessed:01/04/2025

[26] Salemi, A. and Zamani, H., 2024, July. Evaluating retrieval quality in retrieval-augmented generation. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2395-2400).

[27] Es, S., James, J., Anke, L.E. and Schockaert, S., 2024, March. Ragas: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150-158).

[28] Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36, pp.46595-46623.

[29] Feng, F., Yang, Y., Cer, D., Arivazhagan, N. and Wang, W., 2020. Language-agnostic BERT sentence embedding.]arXiv preprint arXiv:2007.01852.

[30] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y., 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675..

[31] Ma, Y., Shao, Y., Wu, Y., Liu, Y., Zhang, R., Zhang, M. and Ma, S., 2021, July. LeCaRD: a legal case retrieval dataset for Chinese law system. In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval (pp. 2342-2348).

Bitcoin Price Prediction Using Classical Machine Learning and Deep Learning

Martina Jovanovikj
Faculty of Electrical Engineering and Information Technologies
Ss. Cyril and Methodius University
in Skopje, North Macedonia
martina.jovanovik@gmail.com

Hristijan Gjoreski
Faculty of Electrical Engineering and Information Technologies
Ss. Cyril and Methodius University
in Skopje, North Macedonia
hristijang@feit.ukim.edu.mk

***Abstract* — This paper explores the application of both classical machine learning (ML) models and deep learning (DL) models for predicting Bitcoin's closing price. Given the highly volatile and non-linear nature of cryptocurrency markets, accurate price prediction poses a significant challenge. We evaluated six classical ML models, including K-Nearest Neighbors (KNN), Linear Regression, LASSO Regression, Support Vector Regression (SVR), Gradient Boosting Regression (GBR), and XGBoost, to assess their effectiveness in forecasting Bitcoin's closing prices. In addition, we investigate deep learning approaches, particularly Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Gated Recurrent Unit and Long Short-Term Memory networks (LSTM), which have shown promise in time series forecasting. Each model's performance is assessed using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), providing a comprehensive comparison of their predictive abilities. Results indicate that while deep learning models, particularly GRU and LSTM, demonstrate potential in capturing long-term dependencies and trends in the data, LASSO Regression stands out as the top-performing model. It outperforms all deep learning models in terms of accuracy, achieving a mean absolute error (MAE) of 849 USD, while the best-performing deep learning model, GRU, recorded an MAE of 913 USD. This study contributes to the growing body of research on cryptocurrency price prediction by offering insights into the relative strengths and limitations of classical machine learning and deep learning models in financial forecasting.**

Keywords — Bitcoin price prediction, machine learning, deep learning, time series forecasting.

I. Introduction

The cryptocurrency market, with Bitcoin as its flagship, has gained significant attention in recent years due to its decentralized nature, potential for high returns, and inherent volatility. Bitcoin [1], the first and most widely known cryptocurrency, has experienced dramatic price fluctuations since its inception. Predicting its closing price is a complex task due to the market's speculative nature, the influence of external factors such as regulations, news, and investor sentiment, as well as the intricate dynamics of the underlying block chain technology. As such, accurate price forecasting holds substantial value for investors, traders, and financial institutions involved in the cryptocurrency ecosystem.

Over the years, various forecasting models have been proposed to predict the price of Bitcoin [1] [12]. Traditional statistical models, such as time series analysis, have been employed to model Bitcoin's price [17]. However, as the complexity of cryptocurrency markets increases, there is a growing need for more advanced predictive models capable of capturing non-linear patterns and long-term dependencies in the data.

Machine learning (ML) models, particularly classical ones, such as K-Nearest Neighbors (KNN), Linear Regression, LASSO Regression, Support Vector Regression (SVR), Gradient Boosting Regression (GBR), and XGBoost, have proven successful in many prediction tasks, including financial forecasting. These models can effectively learn from historical price data and make predictions based on learned patterns. However, their performance can be limited by the inherent limitations of traditional feature extraction and the inability to capture long-range dependencies in sequential data.

On the other hand, deep learning (DL) models, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, have shown great promise in time series forecasting. LSTM, in particular, is well-suited for sequence-based predictions due to its ability to capture long-term dependencies in temporal data. CNN, though primarily used in image-related tasks, has also been adapted for time series forecasting, showing strong performance in extracting hierarchical features from sequential data.

This paper seeks to explore and compare the performance of classical machine learning models and deep learning architectures in predicting Bitcoin's closing price. By evaluating the accuracy, efficiency, and predictive capabilities of models like KNN, Linear Regression, LASSO Regression, SVR, GBR, XGBoost, CNN, RNN, GRU and LSTM, we aim to provide valuable insights into the effectiveness of each model in handling the challenges posed by the volatile cryptocurrency market.

II. Dataset

The dataset used for training and evaluation consists of Bitcoin price data, specifically focusing on the closing price, obtained from Yahoo Finance for the period 2014 to 2025 [17]. Fig. 1 shows Bitcoin's closing prices along with the 30-day rolling volatility, providing insights into the price trend and market volatility, while Fig. 2 shows the annual average close price of Bitcoin in USD, resampled from daily data. The plot illustrates the yearly price trend, helping to identify long-term patterns and fluctuations over time.

The dataset provides valuable information on Bitcoin's market performance, allowing for in-depth analysis of its price movements, volatility, and trends over the specified period. It includes key market variables for Bitcoin, specifically focused on the BTC-USD ticker (Bitcoin to

USD). The dataset captures daily trading activity and includes the following columns:

- **Date**: The date for each entry, representing a single trading day.
- **Close**: The closing price of Bitcoin in USD at the end of each trading day. This is the most common value used for analysis and prediction tasks.
- **High**: The highest price Bitcoin reached during the trading day.
- **Low**: The lowest price Bitcoin reached during the trading day.
- **Open**: The opening price of Bitcoin at the start of the trading day.
- **Volume**: The total trading volume of Bitcoin in USD during that day, representing the amount of Bitcoin traded across all exchanges.

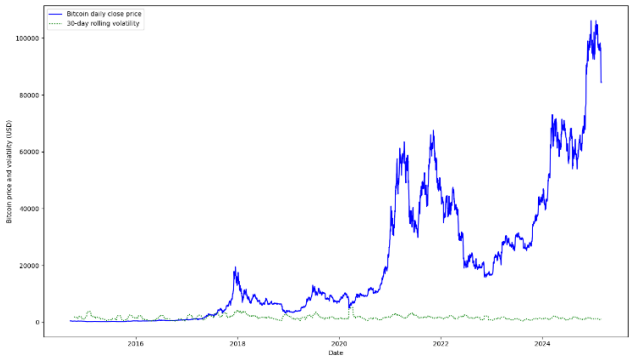


Fig. 1. Daily Bitcoin closing prices (USD) and 30-day rolling volatility over the period from 2014 and 2025

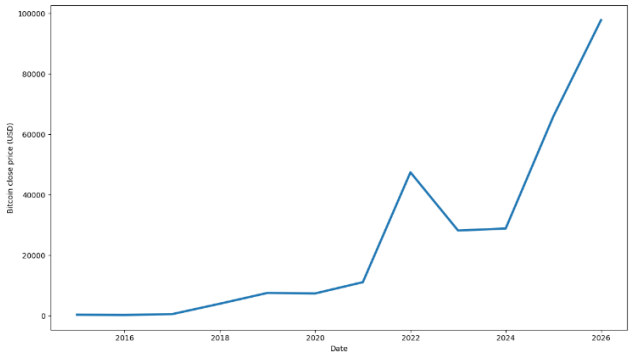


Fig. 2. Resampled yearly Bitcoin closing prices (mean USD) over the period from 2014 to 2025

Table 1 shows the descriptive statistics of Bitcoin's close price, which include the mean, standard deviation, and various percentiles (25%, 50%, and 75%). The mean represents the average closing price over the observed period, providing a central tendency for the data. The standard deviation reflects the level of volatility or price variability, with higher values indicating larger fluctuations in the price. The 50% value, also known as the median, represents the middle point of the data, meaning that 50% of the closing prices lie above and 50% lie below it. The 25% and 75% values, or the first and third quartiles, mark the points where 25% and 75% of the data points fall, respectively. These percentiles help to understand the distribution and spread of Bitcoin's close price, providing insights into its range and overall behavior over time.

TABLE I. DESCRIPTIVE STATISTICS OF BITCOIN CLOSE PRICE (USD)

Statistic	**Bitcoin close price (USD)**
Mean	21,233.49
Standard deviation	24,474.48
Minimum	211.73
25th Percentile	1,328.22
Median (50%)	9,701.53
75th Percentile	34,561.13
Maximum	109,114.88

The data is preprocessed by:

- Handling missing values: Any missing data points are dropped.
- Feature scaling: For classical machine learning (ML) models, the closing price data is scaled using StandardScaler, which standardizes the features by removing the mean and scaling to unit variance. For deep learning (DL) models, the closing price data is scaled using MinMaxScaler, which normalizes the data to a range of 0 to 1. This is particularly useful for deep learning models to improve convergence and stability during training.
- Sliding window technique: A time step of 21 days is used to create the feature set for the models. The past 21 days of prices are used to predict the price for the next day.

We did not explicitly perform trend decomposition, seasonality removal, or stationarity transformations such as differencing. While such preprocessing steps are often beneficial, particularly for linear models, we observed that our models – especially LASSO Regression and deep learning architectures such as LSTM and GRU – were able to learn underlying temporal patterns effectively without manual detrending. Nonetheless, we acknowledge that a formal stationarity check (e.g., Augmented Dickey-Fuller test) and experiments with detrended data could offer further insight and are valuable directions for future work.

III. METHODS

This section outlines the methodologies employed to predict Bitcoin's closing price using both classical machine learning (ML) and deep learning (DL) models.

A. Classical Machine Learning Models

Various classical machine learning (ML) algorithms are employed to predict Bitcoin's closing price. These models are often simpler and require less computational power compared to deep learning methods. However, they still capture underlying patterns and trends in the data, making them suitable for comparison with more complex models. In this study, several classical regression techniques are evaluated, each with unique strengths and weaknesses for time series prediction. The models include K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Linear Regression,

LASSO Regression, Gradient Boosting Regressor (GBR), and XGBoost.

K-Nearest Neighbors [14] is a non-parametric method that predicts the value by averaging the closest K training data points to a given test point. The number of neighbors (K) is tuned to optimize performance. KNN works well for capturing local patterns, but can be computationally expensive with large datasets, as it requires calculating distances to all data points.

Support Vector Regression [10] is a regression version of the Support Vector Machine (SVM) that tries to fit the best line within a margin of tolerance, using a kernel trick to handle nonlinear data. SVR handles high-dimensional data well but can be computationally intensive and requires careful tuning of the kernel and parameters.

Linear Regression [14] is a simple method that assumes a linear relationship between the independent variables (lagged closing prices) and the dependent variable (next day's closing price). It is effective for simple relationships but struggles with more complex patterns or interactions between variables.

LASSO Regression [14] is a regularized version of linear regression that adds an L1 penalty to the loss function to prevent overfitting by encouraging sparsity in the model's coefficients. It performs well with many features but can underperform if the relationships in the data are not sparse. LASSO regression can effectively select a subset of important features, making it useful for high-dimensional datasets.

Gradient Boosting Regression [5] is an ensemble method that combines multiple weak learners (typically decision trees) to form a strong predictive model, where each subsequent model corrects the errors made by the previous one. GBR captures complex patterns but is prone to overfitting and can be computationally expensive. GBR generally provides high predictive accuracy but requires careful tuning to avoid overfitting, especially when the dataset is noisy.

XGBoost [10] is a highly efficient implementation of gradient boosting that includes optimizations such as regularization, missing value handling, and parallelization. XGBoost is fast and accurate but requires careful parameter tuning to avoid overfitting. It is widely used in Kaggle competitions and real-world applications due to its robustness and scalability.

B. Deep Learning Models

Deep learning models, known for their ability to model complex relationships in sequential data, are employed to predict Bitcoin's closing price. These models are particularly powerful for time series forecasting, as they can learn from large datasets and capture intricate patterns over time. The deep learning techniques explored in this study include Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU).

Recurrent Neural Networks (RNNs) [9] are a class of neural networks designed for processing sequential data. RNNs use feedback loops to maintain a memory of previous time steps, making them suitable for time series prediction. The key strength of RNNs lies in their ability to capture dependencies in time-ordered data. However, standard RNNs suffer from issues like vanishing and exploding gradients, which can hinder their ability to learn long-term dependencies.

Convolutional Neural Networks (CNNs) [9] are primarily used for image processing, they have also shown promise in time series forecasting. CNNs apply convolutional filters to data sequences to automatically extract features. In the context of Bitcoin price prediction, CNNs can capture local patterns in the data, such as trends over short periods. When applied to time series, CNNs can be used to detect patterns in the data that may not be immediately obvious through traditional feature engineering methods.

Long Short-Term Memory (LSTM) [13] networks are a special type of RNN designed to address the vanishing gradient problem, which allows them to capture long-range dependencies in sequential data. LSTMs include memory cells that store information over long periods and gates that control the flow of information into and out of the memory. This capability makes LSTMs highly effective for predicting time series data, such as Bitcoin's closing price, where long-term dependencies play a critical role in forecasting future trends. LSTMs are often the go-to model for time series prediction tasks due to their ability to effectively learn temporal relationships in complex datasets.

Gated Recurrent Units (GRUs) [10] are a variant of LSTMs that simplify the architecture by combining the forget and input gates into a single update gate. Like LSTMs, GRUs are designed to capture long-range dependencies in sequential data. GRUs are often more computationally efficient than LSTMs while maintaining comparable performance. Due to their simpler structure, GRUs are faster to train and often outperform LSTMs on smaller datasets or with limited computational resources.

IV. EXPERIMENTAL SETUP

A. Evaluation Metrics

The performance of each model is evaluated using three key metrics:

1. Mean Absolute Error (MAE): Measures the average magnitude of the errors in a set of predictions, without considering their direction.
2. Root Mean Squared Error (RMSE): Measures the square root of the average of the squared differences between predicted and actual values, which penalizes larger errors more heavily.
3. Mean Absolute Percentage Error (MAPE) measures the average absolute difference between predicted and actual values, expressed as a percentage of the actual values.

$$MAE = \frac{1}{q}\sum_{1}^{q}\left|BC_{predicted} - BC_{true}\right| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{q}\sum_{1}^{q}\left(BC_{predicted} - BC_{true}\right)^2} \quad (2)$$

$$MAPE = \frac{100}{q}\sum_{1}^{q}\left|\frac{BC_{predicted} - BC_{true}}{BC_{true}}\right| \quad (3)$$

Where q is the number of data sample, $BC_{predicted}$ is the Bitcoin close value predicted by the model, while BC_{true} is the real Bitcoin close value.

These metrics are used to compare the accuracy of all models on the test set, with lower values indicating better model performance.

Additionally, a naive benchmark is used as a baseline, where each day's closing value is predicted to be equal to the previous day's closing value. The MAE, RMSE, and MAPE of each model are compared against this benchmark to assess their relative effectiveness.

B. *Hyperparameter Tuning*

To improve model performance, hyperparameters are optimized using GridSearchCV and RandomizedSearchCV for the classical ML models. For deep learning models, tuning is performed by adjusting the number of layers, units, and the learning rate.

C. *Model Training and Testing*

The data is split into 70% training and 30% testing, with the time component considered. The first 70% of the dataset, representing earlier data, is used for training, and the remaining 30%, representing the most recent data, is reserved for testing. Each model is trained on the training set, and predictions are made on the test set. The results of these predictions are reported, demonstrating each model's performance in forecasting Bitcoin's future closing prices.

V. EXPERIMENTAL RESULTS

Among the classical models, Linear Regression and LASSO Regression demonstrated the best results, exhibiting the lowest Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) compared to other machine learning models. These models were able to provide relatively accurate predictions of the Bitcoin closing price, with Linear Regression showing a slight edge in raw prediction accuracy, while LASSO Regression offered greater robustness across different metrics.

Notably, the naive benchmark also performed competitively, achieving similar or better results than most complex models. This highlights the inherent difficulty of the prediction task and suggests that some models may struggle to outperform a simple baseline. LASSO Regression was the only model that consistently matched or slightly outperformed the benchmark across all evaluation metrics, underscoring its effectiveness despite its simplicity.

On the other hand, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Gradient Boosting Regressor (GBR), and XGBoost exhibited significantly higher error metrics, indicating weaker predictive performance on this dataset. This suggests that these models may be more sensitive to noise or less suited for the time series characteristics of Bitcoin prices.

Fig. 3 shows a comparison of actual and predicted data using the LASSO Regressor, while Fig. 4 illustrates the evaluation metrics across the classical models, including the naive benchmark.

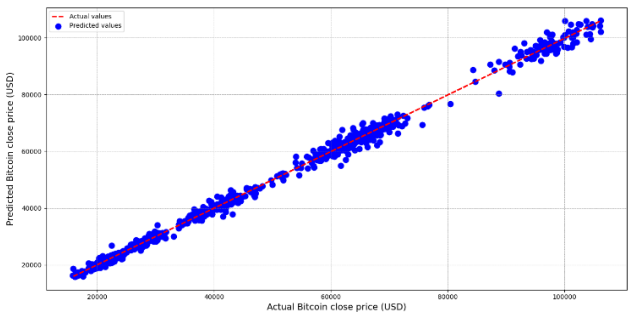


Fig. 3. Actual vs predicted Bitcoin close values (USD) using the best performing classical ML algorithm – LASSO Regressor

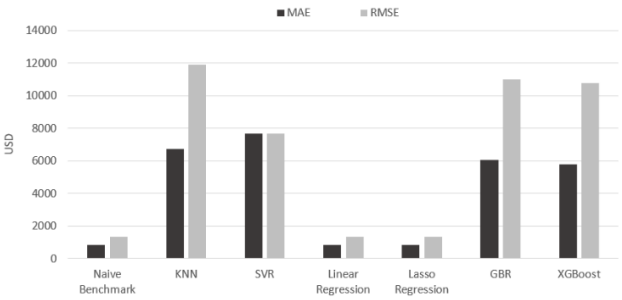


Fig. 4. Comparison of the MAE and RMSE for the classical ML algorithms

In contrast to the classical models, deep learning approaches such as GRU and LSTM also delivered competitive results, demonstrating their potential to model complex temporal dependencies in Bitcoin price data. Among them, the GRU model performed particularly well, achieving a MAPE close to that of the naive benchmark. This suggests that GRU was effective at learning sequential patterns, offering a strong balance between accuracy and model simplicity.

LSTM showed similar performance in terms of MAE and RMSE, but its significantly higher MAPE indicated that the model may have experienced instability or convergence issues, possibly due to overfitting, sensitivity to hyperparameters, or limitations in the available dataset. These discrepancies highlight the challenges in training deep neural networks, where even models with similar loss values may behave differently in practical forecasting scenarios.

Other architectures, including CNN and RNN, underperformed both GRU and the naive benchmark, reflecting limited ability to generalize in this specific financial time series task. This underperformance may be attributed to architectural mismatches with the nature of the data or insufficient tuning of convolutional and recurrent layers.

Overall, these findings reinforce the strength of the naive baseline in short-term Bitcoin forecasting and emphasize the need for thoughtful model design, regularization, and robust feature engineering when applying deep learning to volatile and noisy financial datasets. They also suggest that simpler architectures like GRU can outperform deeper or more complex models when appropriately tuned. Fig. 5 shows the actual vs. predicted Bitcoin close prices using GRU, while Fig. 6 presents the evaluation of all deep learning models, including the benchmark.

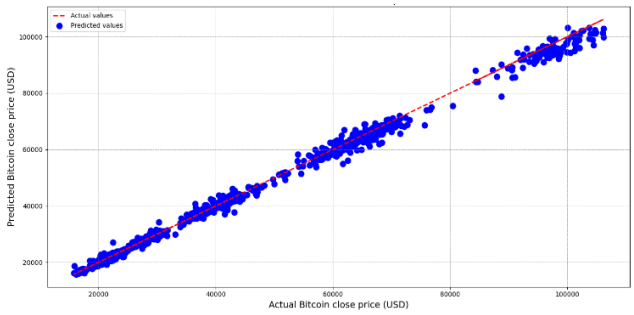


Fig. 5. Actual vs predicted Bitcoin close values (USD) using the best performing DL algorithm – GRU

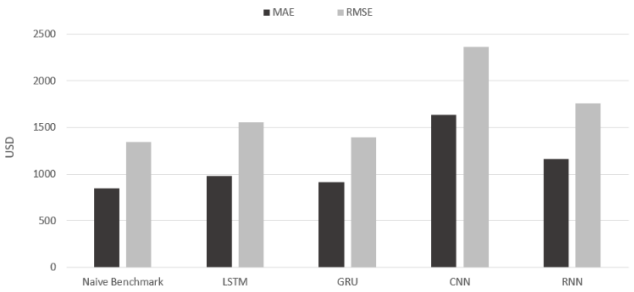


Fig. 6. Comparison of the MAE and RMSE for the DL algorithms

Fig. 7 shows the comparison of the MAPE for both the classical ML and DL algorithms including naive benchmark. Fig. 8 shows the test vs predicted Bitcoin close values (USD) using the best performing classical ML model – LASSO Regressor, while Fig. 9 shows the test vs predicted Bitcoin close values (USD) using the best performing DL model – GRU.

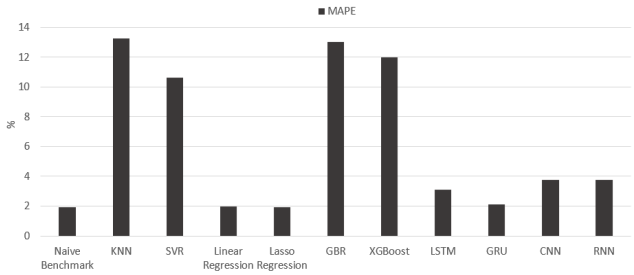


Fig. 7. Comparison of the MAPE for both the classical ML and DL algorithms

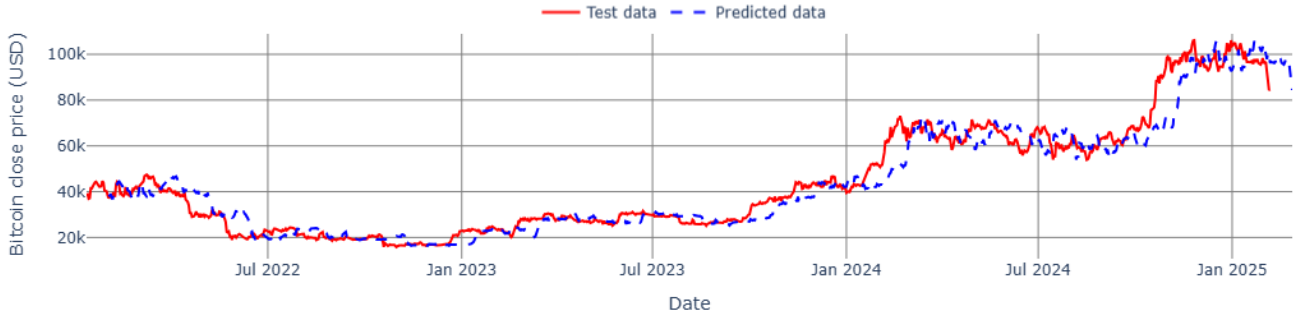


Fig. 8. Bitcoin close price – test vs predicted values using the best performing classical ML algorithm – LASSO Regressor

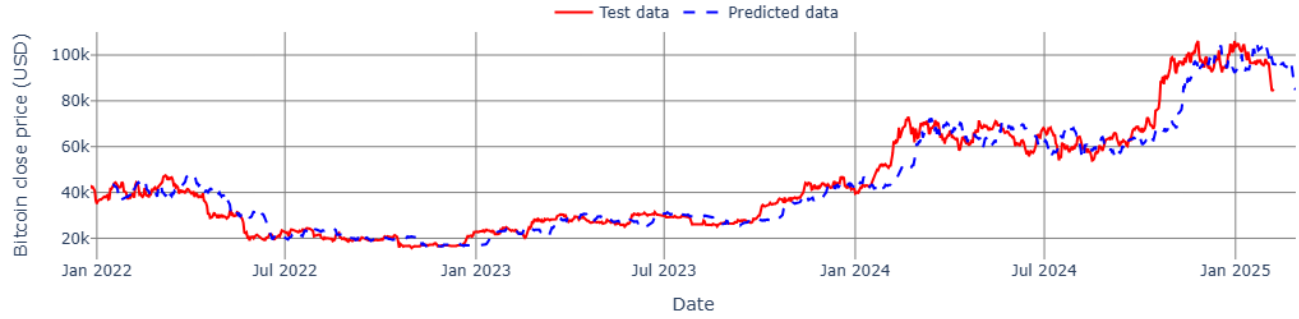


Fig. 9. Bitcoin close price – test vs predicted values using using the best performing DL algorithm – GRU

VI. CONCLUSIONS

In this study, we compared multiple machine learning (ML) and deep learning (DL) models to predict Bitcoin's closing price. The performance of the models was evaluated based on three key metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

Among the Machine Learning (ML) models, LASSO Regression delivered the most reliable results, achieving the lowest MAE, RMSE, and MAPE values. This suggests that LASSO was able to effectively capture the underlying trends in the Bitcoin price time series while maintaining simplicity and robustness. Even when compared to more complex models like XGBoost and Gradient Boosting Regressor, LASSO maintained a clear performance edge. Notably, LASSO was the only classical model that consistently outperformed the naive benchmark, reinforcing its practical value in financial forecasting.

In the Deep Learning (DL) category, LSTM and GRU outperformed other architectures, with GRU slightly leading in overall performance. Both models excelled in modeling the sequential dependencies of Bitcoin's price, demonstrating strong predictive power, especially for capturing medium- to long-term patterns. However, GRU's performance was more stable across all evaluation metrics, including MAPE, indicating better generalization.

In contrast, CNN and RNN underperformed, with RNN slightly ahead of CNN but still falling behind LSTM and GRU. Their limited ability to outperform the naive baseline suggests that they may not be well-suited for this specific forecasting task without further architectural adjustments or feature enhancements.

The strong performance of the naive benchmark across many models underscores the inherent difficulty of Bitcoin price prediction and the volatility of financial time series data. It also emphasizes the importance of including such a

baseline in future studies, as it provides a realistic standard for model evaluation.

Overall, the results indicate that carefully regularized models like LASSO and well-tuned recurrent architectures like GRU offer promising directions for short-term Bitcoin forecasting, balancing accuracy and model complexity effectively.

REFERENCES

[1] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System", Oct. 2008 [Online]. Available: https://bitcoin.org/bitcoin.pdf.

[2] P. Sudhakaran, V. Sharma, and S. Khandelwal, "Bitcoin Price Prediction Using Recurrent Neural Networks and Long Short-Term Memory," E3S Web of Conferences, vol. ICECS'24, Apr. 2024.

[3] M. Frohmann, M. Karner, S. Khudoyan, R. Wagner, M. Schedl "Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting," Big Data Cognition Comput., vol. 7, no. 3, Jul. 2023.

[4] P. L. Seabe, C. R. B. Moutsinga, E. Pindza, "Forecasting Cryptocurrency Prices Using LSTM, GRU, and Bi-Directional LSTM: A Deep Learning Approach," Fractal Fract., vol. 7, no. 2, pp. 203-203, Feb. 2023.

[5] N.S.S. K. Balijepalli, V. Thangaraj, "Prediction of cryptocurrency's price using ensemble machine learning algorithms", European Journal of Management and Business Economics, vol. 34, no. 1, Jan. 2025

[6] A. Ladhari, H. Boubaker, "Deep Learning Models for Bitcoin Prediction Using Hybrid Approaches with Gradient-Specific Optimization," MDPI Journals, vol. 6, no. 2, Apr. 2024.

[7] Z. Shengao, L. Mengze, and Y. Chunxiao, "The Empirical Analysis of Bitcoin Price Prediction Based on Deep Learning Integration Method", Forecasting, vol. 6, no. 3, Aug. 2024.

[8] S. Erfanian, Y. Zhou, A. Razzaq, A. Abbas, A. A. Safeer, T. Li, "Investigating the Problem of Cryptocurrency Price Prediction: A Deep Learning Approach", PMC, May 2020.

[9] S. Alonso-Monsalve, Andrés L. Suarez-Certulo, A. Cervantes, D. Quintana, "Convolution on Neural Networks for High-Frequency Trend Prediction of Cryptocurrency Exchange Rates", CERN Publication, Jan. 2024.

[10] D. L. Johnson, S. Binnewies, B. Stantic, "Cryptocurrency Price Prediction Algorithms: A Survey and Future Directions", MDPI Journals, vol. 6, no. 3, Aug. 2024.

[11] J. Chen, "Analysis of Bitcoin Price Prediction Using Machine Learning" Journal of Risk and Financial Management, vol. 16, no. 1, Jan. 2023.

[12] A. Alkamali, "Bitcoin Short-term Price Prediction Using Time Series Analysis", Rochester Institute of Technology Theses, Feb. 2024.

[13] K. Davies, J. Liu, and Y. Zhang, "Bitcoin Price Prediction Using LSTM and Technical Indicators", Journal of Intelligent Information Systems, vol. 57, no. 2, pp. 257–273, Oct. 2022.

[14] L. S. Reddy, Dr.P. Sriramya, "A Research On Bitcoin Price Prediction Using Machine Learning Algorithms", International Journal of Scientific & Technology Research, vol. 9, no. 04, April 2020..

[15] S. Cenoy, K.Pathak, Maneesha, "Predictive Analytics for Cryptocurrency Prices Using Machine Learning", Journal of Electrical Systems, Vol. 20 No. 10s, pp. 1964 – 1971, Oct. 2024.

[16] S. Penmetsa, M. Vermula, "Cryptocurrency Price Prediction with LSTM and Transformer Models Leveraging Momentum and Volatility Technical Indicators", IEE, vol. ICDSCA'23, Oct. 2023.

[17] Yahoo Finance, "Bitcoin Historical Data (BTC-USD)", Yahoo Finance, Mar. 2025. [Online]. Available: https://finance.yahoo.com/quote/BTC-USD/history.

Quantitative Analysis of Tunnel Stability Using a Data-Driven Approach

1st Mite Mazgaliev
Faculty of computer scince and engineering
ss. Cyril and Methodius University
Skopje, North Macedonia
mite.mazgaliev@students.finki.ukim.mk

2nd Stefan Andonov
Faculty of computer scince and engineering
ss. Cyril and Methodius University
Skopje, North Macedonia
stefan.andonov@finki.ukim.mk

3rd Bojan Ilijoski
Faculty of computer scince and engineering
ss. Cyril and Methodius University
Skopje, North Macedonia
bojan.ilijoski@finki.ukim.mk
https://orcid.org/0000-0002-5422-7353

4th Zlatko Zafirovski
Faculty of civil engineering
ss. Cyril and Methodius University
Skopje, North Macedonia
zlatko.zafirovski@gf.ukim.edu.mk

***Abstract*—This study analyzes key geotechnical parameters—Rock Mass Rating (RMR), Quality Index (Q), and Deformation Modulus (D)—alongside potential predictors. The research defines these parameters, outlines measurement techniques, and examines data collection methods. Using existing geotechnical data from various locations, statistical analyses are conducted to identify significant correlations between key parameters and their predictors. The findings provide insights into the factors influencing rock mass behavior and deformation, contributing to improved geotechnical assessments and engineering applications.**

***Index Terms*—Rock Mass Rating (RMR), Quality Index (Q), Deformation Modulus (D), Statistical analysis, Correlation analysis, Engineering applications**

I. Introduction

The measurement of key geotechnical parameters—Deformation Modulus (D), Quality Index (Q), and Rock Mass Rating (RMR)—is essential for geotechnical analysis and design. However, their direct measurement is often expensive, time-consuming, and complex. This challenge highlights the need to explore alternative methods for estimating these parameters.

This study investigates the relationships between key parameters and more easily measurable geotechnical properties. By analyzing existing data and applying statistical techniques, the goal is to identify significant correlations that can provide a reliable way to estimate D, Q, and RMR through simpler, cost-effective measurements. This approach could improve efficiency in geotechnical assessments while maintaining accuracy in rock mass characterization.

The data analysis and visualizations in this research were done in Python [1], using visualization packages [2].

II. Related Work

Direct measurement of the Deformation Modulus is both costly and time-consuming. Consequently, several researchers have developed estimation models to approximate this parameter. Early efforts by [3] introduced an empirical model for estimation. Subsequent studies, such as those by [4] and [5], expanded on these ideas, with [5] proposing a non-linear equation for estimation.

Additionally, [6] performed a regression analysis using Rock Mass Rating (RMR) and Geological Strength Index (GSI) to approximate the Deformation Modulus; however, the resulting models were statistically insignificant for accurate prediction. More recently, [7] implemented several empirical approximations based on a dataset of 115 measurements, yielding estimations that closely matched the observed values.

III. Analysis

In this section, we will go through all the details and analysis regarding the parameters and the data.

A. Data

The data used in the research was provided by UKIM - Faculty of Civil Engineering; there are 109 measurements made of different sites for various projects across the country. The measured parameters were:

- Overlay (m)
- Volumetric weight (kN/m³)
- Seismic wave speed Vp
- Compressive strength Mpa
- Water inflow (l/min)
- Rock quality designation RQD
- Mean distance between cracks (mm)
- Poisson Coefficient

For each measurement there is the calculated and measured grade for quality for the rock masses.

- RMR - Rock mass rating
- Quality Index Q
- Deformation Modulus D

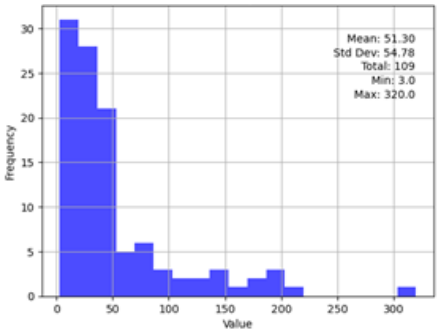


Fig. 1. Overlay summary statistics data visualization

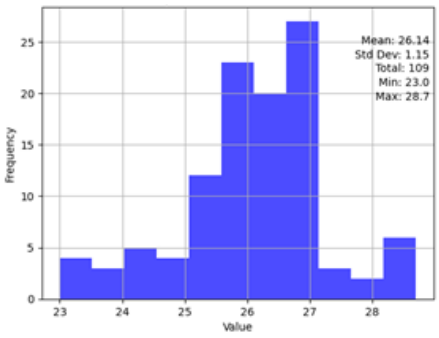


Fig. 2. Volumetric weight summary statistics data visualization

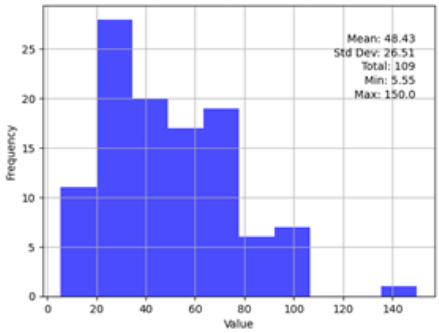


Fig. 3. Compressive strength summary statistics data visualization

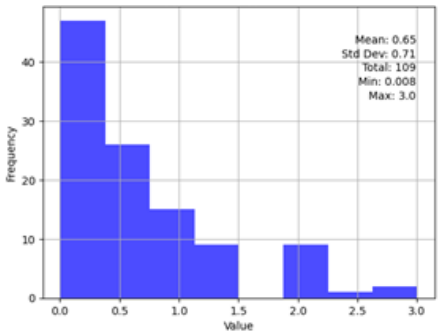


Fig. 4. Water inflow summary statistics data visualization

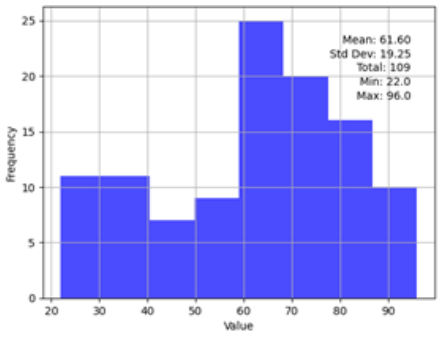


Fig. 5. RQD summary statistics data visualization

B. Definitions

1) Overlay (M): The rock overlay fig[1], the rock layers that lie above a geological feature or rock mass of interest. This overlying material can significantly influence the mechanical behavior of the underlying rock by affecting stress distribution, deformation patterns, and overall stability.

2) Volumetric weight (kN/m³): Volumetric weight fig[2], also known as specific weight, is defined as the weight of a material per unit volume. Expressed in kilonewtons per cubic meter (kN/m³), this parameter is critical in geotechnical engineering as it influences stress distribution and overall stability in soil and rock masses.

3) Compressive strength Mpa: Compressive strength fig[3] is measured along a single axis and represents the rock's ability to resist deformation and failure under compression. This property is crucial in geotechnical engineering for assessing the stability and load-bearing capacity of rock structures.

4) Water inflow (l/min): Water inflow fig[4] is a critical parameter in tunnel and mine excavations, as it significantly affects the stability of underground structures. It represents the volume of water entering an excavation per unit of time, typically measured in liters per minute (l/min) over a standardized excavation length, such as 10 meters of tunnel advance.

5) Rock quality designation RQD: RQD fig[5], or the degree of deformations in a rock mass, is measured by taking a few cores and measuring the small deformations that are equal to or greater than 100mm. And the result is calculated with:

$$RQD = \frac{N}{M} \tag{1}$$

where:

- N = The sum of the length of core deformations that are **100 mm or longer**.
- M = The sum of the length of cores examined.

6) Mean distance between cracks (mm): The mean distance between deformations, and in this case every discontinuity is counted as a crack; faults, foliation, layering

7) Seismic wave speed Vp: Seismic waves fig[7] are a type of mechanical waves of acoustic energy which passes trough

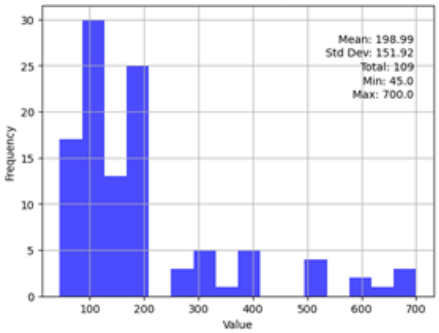


Fig. 6. Mean distance between cracks summary statistics data visualization

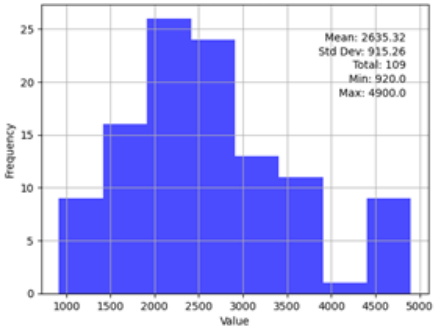


Fig. 7. Seismic wave speed summary statistics data

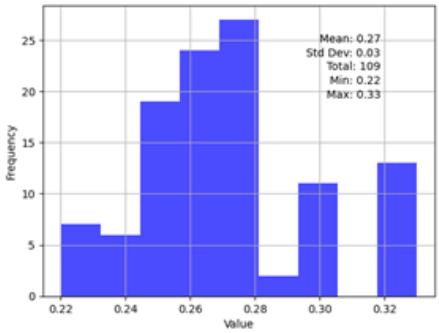


Fig. 8. Poisson coefficient summary statistics data

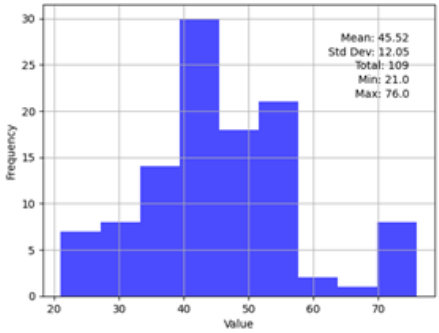


Fig. 9. RMR summary statistics data

the ground and other planetary masses. These waves can come from eruptions from a volcano, massive land subsidence, earthquakes and explosives.

$$V_p = \sqrt{\frac{K + \frac{4}{3}G}{\rho}} \tag{2}$$

where:

- V_p Is the speed of the compressive seismic waves passing through the material in m/s.
- K Is a measurement for the resistance of the material under uniform compression.
- G Is a measure for resistance to shear transformation.
- ρ is the density of the material.

8) Poisson Coefficient: It is the quotient of the lateral stress to the axial stress, most rigid masses have a Poisson's ratio of about 0.3. The Poisson coefficient fig[8] tells us how a rock mass will react when we give it lateral and axial stress with which we get the change.

$$\text{Poisson Coefficient } = |\frac{-\epsilon_l}{\epsilon_a}| \tag{3}$$

where:

- ϵ_l is the lateral strain.
- ϵ_a is the axial strain.

The way this is tested is by taking a cylindrical part from the rock mass, and is subjected to compression tests ex. large axial stress.

9) Rock Mass rating - RMR: Is a system for classification of the rock mass utilizing the following fig[9] parameters:

- Compressive strength
- RQD - Rock quality designation.
- Mean distance between cracks.
- Crack condition, which refers to the opening, the type of filler and its thickness, continuity, and roughness.
- The state of the water inflow or water pressure which is very important for the stability of the underground excavation..
- The sixth criterion is considered additional and refers to the orientation of the most relevant fracture system, in relation to the orientation of the excavation.

The range for the values for RMR is 0-100, or classified from 'Very bad rock mass' to 'Very good rock mass' based on the [8] table 12.1.E.

With these measurements we can figure out the:

- The stable time standing of the underground opening.
- The type of support required for the opening.
- The length excavation progress.

10) Quality Index Q: Similar to RMR the quality index fig[10] is given 6 classes.:

- RQD
- Index relating to the number of crack systems Jn
- Parameter relating to the roughness of the cracks Jr
- Parameter relating to the variation of the cracks Ja
- Index for reduction due to water pressure Jw

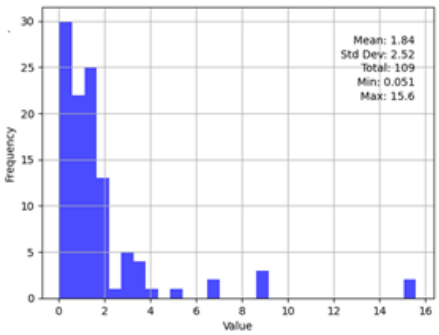


Fig. 10. Quality index summary statistics data

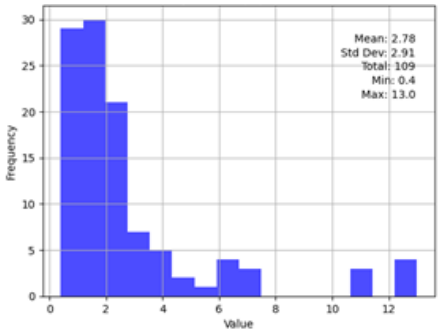


Fig. 11. Deformation modulus summary statistics data

- SRF, stress reduction factor

And is calculated using:

$$Q = \frac{RQD}{J_n} \times \frac{J_r}{J_a} \times \frac{J_w}{SRF} \tag{4}$$

The first factor is related to the geoemtry of the rock mass, the size and shape. Q grows if RQD grows and falls when Jn falls. The second factor represents the roughness and the friction properties of the walls or fillers. The last factor shows the stress state in the mass.

11) Deformation Modulus D: The deformation modulus fig[11] is one of the best parameters to explain the mechanical ratio between the rock and rock masses, especially in underground excavation. [9]

There are a few ways on sitetests to get the Deformation modulus:

- The Goodman Jack test.
- Plate loading test.

The theoretical measurement of the Deformation modulus is very time consuming and expensive, and is often approximated using RMR or Quality index or other values. Example approximations:

$$Em = 2 * RMR - 100, \text{ if } RMR > 50 \text{ [10]} \tag{5}$$

Which is good for measurements where RMR is greater than 50.

$$Em = 10^{\frac{(RMR-10)}{40}}, \text{ if } RMR < 50 \text{ [4]} \tag{6}$$

Which gives us good results for RMR values under 50.

$$E_m = E_i(0.0028RMR^2 + 0.9\exp^{RMR/22.82}) \text{ [5]} \tag{7}$$

$$Em = 7 \pm \sqrt{10((RMR - 44)/21)} \text{ [11]} \tag{8}$$

C. Correlation

In this section we will look over the statistical visualizations and explain the meaning and conclusions gathered

The first observation is that the key parameters are highly correlated with each other. For instance, RMR (Rock Mass Rating) shows a strong correlation with both the Quality Index and Deformation Modulus 12. Similarly, the Quality Index and Deformation Modulus also have a high correlation. These high correlations also explain the already derived formulas to approximate the deformation modulus. The expected high correlation between RMR and the predicting parameters, which are defined in the calculation of the RMR score (such as Water Inflow, RQD (Rock Quality Designation), Mean Distance Between Cracks, and Compressive Strength), is evident.

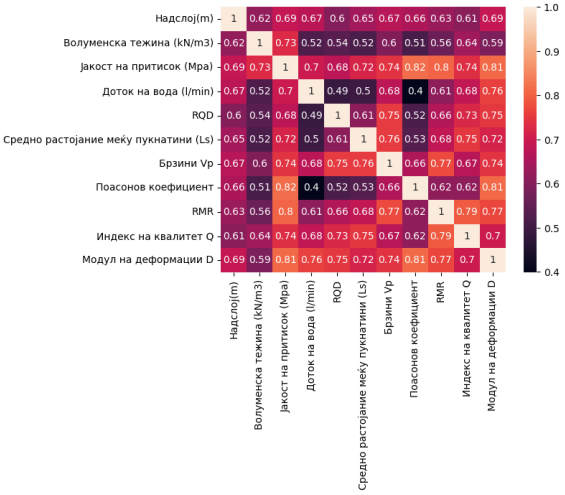


Fig. 12. MIC heat map

Beyond the correlation between the key parameters, the Deformation Modulus exhibits a strong correlation with other factors such as Compressive Strength, Poisson's Ratio, RQD, Water Inflow, Mean Distance Between Cracks, and Seismic Wave Speed. This suggests further examination of these parameters to develop models that predict the Deformation Modulus based on them.

IV. Conclusion and Discussion

Based on our analysis, even with limited data, we were able to extract meaningful correlations between the key geotechnical parameters and their predictors, and our calculated correlation values are in agreement with theoretical expectations. This opens an interesting avenue for future work: developing

predictive models for Rock Mass Rating (RMR), Deformation Modulus, and Quality Index using these predictors.

One promising research direction is to investigate whether RMR and Quality Index can be reliably estimated using a subset of the predictor variables, thereby reducing the complexity and cost of direct measurements without sacrificing significant accuracy. Additionally, it may be feasible to predict the Deformation Modulus solely from these predictors, bypassing the need to directly estimate RMR or Quality Index as intermediate steps. This approach could streamline the assessment process and improve efficiency in geotechnical design.

Future research should focus on validating these preliminary findings with larger datasets, performing sensitivity analyses to identify the most influential predictors, and comparing alternative modeling approaches, including statistical and machine learning techniques. Such studies could lead to robust, practical tools for accurately predicting key geotechnical parameters based on easily measurable variables, ultimately advancing our understanding of rock mass behavior in engineering applications.

ACKNOWLEDGMENT

This work is partially supported by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje

REFERENCES

[1] https://www.python.org/. [Online]. Available: https://www.python.org/
[2] https://matplotlib.org/. [Online]. Available: https://matplotlib.org/
[3] Z. Bieniawski, "Determining rock mass deformability: experience from case histories," in *International journal of rock mechanics and mining sciences & geomechanics abstracts*, vol. 15, no. 5. Elsevier, 1978, pp. 237–247.
[4] J. L. SERAFIM, "Consideration of the geomechanical classification of bieniawski," in *Proc. int. symp. on engineering geology and underground construction*, vol. 1, 1983, pp. 33–44.
[5] G. Nicholson and Z. Bieniawski, "A nonlinear deformation modulus based on rock mass classification," *International journal of Mining and geological engineering*, vol. 8, pp. 181–202, 1990.
[6] A. Kayabasi, C. Gokceoglu, and M. Ercanoglu, "Estimating the deformation modulus of rock masses: a comparative study," *International Journal of Rock Mechanics and Mining Sciences*, vol. 40, no. 1, pp. 55–63, 2003.
[7] C. Gokceoglu, H. Sonmez, and A. Kayabasi, "Predicting the deformation moduli of rock masses," *International Journal of Rock Mechanics and Mining Sciences*, vol. 40, no. 5, pp. 701–710, 2003.
[8] Ö. Aydan, *Rock mechanics and rock engineering: Volume 1: Fundamentals of rock mechanics*. CRC Press, 2019.
[9] A. Palmström and R. Singh, "The deformation modulus of rock masses—comparisons between in situ tests and indirect estimates," *Tunnelling and Underground Space Technology*, vol. 16, no. 2, pp. 115–131, 2001.
[10] Z. Bieniawski, *Tunnel design by rock mass classifications*. The Station, 1979, vol. 79, no. 19.
[11] M. Diederichs and P. Kaiser, "Stability of large excavations in laminated hard rock masses: the voussoir analogue revisited," *International Journal of Rock Mechanics and Mining Sciences*, vol. 36, no. 1, pp. 97–117, 1999.

Automatic Speech Recognition for Swiss German: A Comparative Analysis of Whisper and Wav2Vec

1st Marija Stojcheva
Faculty of Computer Science and Engineering Ss. Cyril and Methodius University
Skopje, N. Macedonia
marija.stojcheva@finki.ukim.mk

2nd Goran Petkovski
Faculty of Computer Science and Engineering Ss. Cyril and Methodius University
Skopje, N. Macedonia
goran.petkovski@finki.ukim.mk

3rd Igor Mishkovski
Faculty of Computer Science and Engineering Ss. Cyril and Methodius University
Skopje, N. Macedonia
igor.mishkovski@finki.ukim.mk

***Abstract*—This survey provides a comprehensive review of the challenges and advancements in Automatic Speech Recognition for dialects, focusing on Swiss German. It compares the performance of two state-of-the-art models, Whisper and Wav2Vec 2.0, in dialectal transcription tasks. While Whisper excels in zero-shot scenarios, its accuracy for dialects remains limited without fine-tuning. In contrast, Wav2Vec 2.0, leveraging self-supervised learning, demonstrates strong performance in low-resource settings with scarce labeled data. The study also examines evolving evaluation metrics, advocating for approaches that capture semantic preservation beyond traditional Word Error Rate and BLEU scores. Key challenges including data scarcity, phonetic variability, and model efficiency are discussed alongside potential solutions such as Parameter-Efficient Fine-Tuning and hybrid models like Whisper-turbo. By addressing these issues, this survey contributes to the development of more inclusive ASR systems capable of handling Swiss German and other low-resource dialects.**

***Index Terms*—automatic speech recognition, dialect transcription, Swiss German, Whisper, Wav2Vec**

I. Introduction

Automatic Speech Recognition (ASR) is a technology that converts spoken language into written text, enabling machine interpretation and processing of human speech. It has become an integral component of various applications, ranging from voice assistants and real-time captioning to automated transcription and customer service systems [1] [2]. The increasing adoption of ASR is largely attributed to advancements in deep learning, which have significantly improved transcription accuracy [3] [4]. These improvements have enhanced accessibility, allowing individuals with hearing impairments to engage with spoken content through captions and transcriptions [5]. ASR has also facilitated more efficient human-computer interaction, enabling hands-free operation of devices, including smart home assistants, voice-controlled navigation systems, and dictation software [6] [7] [8]. Additionally, companies use ASR for customer service automation, optimizing interactions through chatbots and call center transcription, while educators and researchers use it to support language learning and linguistic analysis [9] [10]. In the healthcare domain, ASR is increasingly integrated into the medical field, streamlining clinical documentation, reducing manual note-taking for healthcare professionals, and improving patient interactions [11] [12]. Furthermore, ASR is making significant strides in forensic analysis, including speaker identification and legal transcription for courtroom proceedings [13] [14].

Despite these advancements, ASR technology continues to face several persistent challenges that affect its accuracy and overall performance. A fundamental challenge is the inherent variability of speech signals [15] [16]. Even when the same phrase is spoken by the same individual multiple times, subtle differences in tone, speed, accent, and emotional state introduce variations that complicate accurate recognition. Moreover, unclear phoneme or word boundaries further increase recognition difficulties. External factors, such as background noise and echo, also degrade transcription quality, reducing the reliability of ASR systems in real-world conditions. Another major challenge is speaker differentiation in multi-speaker environments [17]. Speaker diarization, the process of distinguishing and attributing speech to different speakers, is critical for transcription accuracy. However, this task becomes particularly complex in conversations with overlapping speech or acoustically similar speakers, where ASR models struggle to distinguish between different voices. Beyond these general challenges, dialect transcription, especially for low-resource dialects such as Swiss German, introduces further complexities. These dialects often have unique phonetic and grammatical features, lack standardized orthography, and suffer from limited transcribed data.

To address these challenges, this paper provides a comprehensive analysis of the transcription of the Swiss German dialect. It first examines the historical evolution of ASR technology, tracing its progression from traditional Hidden Markov Models to modern deep learning approaches. The paper then delves into the specific obstacles encountered when transcribing Swiss German, focusing on the impact of data scarcity, lack of standardization, and grammatical divergence. Furthermore, it explores the performance of two leading ASR models, OpenAI's Whisper and Meta AI's Wav2Vec 2.0, on Swiss German datasets. Through a detailed analysis of their architectures, training methodologies, and empirical results, we aim to identify the strengths and weaknesses of each model in this challenging domain. This research aims to contribute to the development of more inclusive and robust ASR systems

by addressing the complexities of dialect transcription and promoting greater linguistic inclusion.

II. Evolution and Methodologies in ASR Systems

ASR technology has advanced significantly over the years, evolving from its initial application in recognizing basic commands and digits to becoming a core technology that facilitates human-machine interaction [18]. Early ASR systems introduced in the 1970s, relied on Hidden Markov Models (HMMs), which provided a statistical framework for recognizing speech by modeling sequential patterns. These models utilized a traditional ASR pipeline consisting of a lexicon model for phonetic pronunciations, an acoustic model to map audio signals to speech sounds, and a language model to predict word sequences based on linguistic probabilities. While HMMs enabled advancements such as Large Vocabulary Continuous Speech Recognition (LVCSR), they required extensive feature engineering and struggled with speech variability, such as accents, speaking styles, and noisy environments [19].

The integration of deep learning techniques in the 2010s revolutionized ASR technology, overcoming many limitations of traditional HMM-based systems. Deep Neural Networks (DNNs), particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, enhanced performance by capturing temporal dependencies in speech data [20] [21] [22]. LSTMs addressed the vanishing gradient problem, making them effective at modeling long-range dependencies and improving the system's ability to understand contextual speech.

The development of end-to-end ASR systems further revolutionized the field speech recognition. Deep Speech, directly mapped acoustic features to word sequences, removing the need for separate lexicon and language models [23]. This innovation simplified the ASR pipeline, improving both efficiency and robustness. Following this development, models such as Connectionist Temporal Classification (CTC) and Listen, Attend, and Spell (LAS) refined the end-to-end approach [24] [25]. CTC streamlined the mapping process, although it still relied on external language models for optimal performance. LAS, on the other hand, integrated attention mechanisms, enabling the system to focus on the most relevant segments of speech, which greatly enhanced accuracy in noisy environments. Transducer models (RNNTs) expanded these concepts by combining both acoustic and language models into a single architecture, leading to improved robustness and performance in complex scenarios [26].

The introduction of Transformer models marked another significant advancement in ASR systems [27] [28]. Initially designed for natural language processing, Transformer models utilize self-attention mechanisms that allow the model to capture long-range dependencies and contextual information within speech data. This capability led to substantial improvements in recognition accuracy, particularly in acoustically challenging environments. The hybrid architecture of Conformers, combining Transformers with convolutional layers, proved crucial for capturing both local and global information, leading to improved performance in speech recognition [29].

OpenAI's Whisper and Meta AI's Wav2Vec 2.0 represent significant advancements in ASR, each with distinct approaches. Whisper, a supervised model trained on a massive multilingual dataset, excels in transcribing and translating across 99 languages [30]. In contrast, Wav2Vec 2.0 and its successor XLS-R utilize self-supervised learning to learn from unlabeled audio data, reducing the dependency on labeled data and facilitating cross-lingual transfer learning [31] [32].

III. Challenges in Transcribing Dialects

Despite significant advancements in ASR technology, dialect transcription remains a persistent challenge, particularly for less-standardized and resource-limited dialects [33]. Many dialects lack a standardized writing system, complicating the establishment of consistent transcription conventions. Additionally, dialects often feature unique grammatical structures, such as variations in verb conjugation, word order, and the use of articles and prepositions, which diverge from the corresponding standard language and further increase the complexity of accurate transcription.

Key challenges also include:

- Scarcity of Annotated Data: Many dialects, especially those with strong oral traditions, have limited transcribed speech corpora. The lack of sufficient labeled training data prevents ASR models from effectively learning the linguistic patterns unique to these dialects.
- Phonetic and Lexical Variability: Dialects often exhibit significant phonetic variation within small geographic areas. For example, in some German dialects, the Standard German word "Apfel" (apple) may be pronounced as "Apfel" in one region but as "Äpfel" or "Appel" in others, reflecting notable phonetic shifts.
- Code-Switching and Code-Mixing: In multilingual societies, speakers frequently switch between languages within a single sentence. In India, for instance, Hinglish (Hindi-English code-mixing) is common, requiring ASR systems to handle multiple linguistic structures simultaneously.
- Domain-Specific Limitations: Although some dialectal ASR models perform well in controlled environments, such as news broadcasts or parliamentary debates, they often struggle with spontaneous, conversational speech. Spontaneous speech typically includes disfluencies, hesitations, and rapid speech, further complicating transcription accuracy.

This issue is not limited to a single region but is a global challenge. For example, Arabic's vast geographical distribution and dialectal variation complicate transcription. The MGB-5 dataset contains over 3,000 hours of dialectal Arabic speech from 17 countries, capturing the differences in pronunciation and grammar across the Middle East and North Africa [34]. Similarly, the MUCS 2021 dataset tackles challenges in the multilingual and code-switching context of the Indian subcontinent. It includes tasks for building an ASR system for

six Indian languages and handling code-switching between Hindi-English and Bengali-English pairs [35]. In China, the tonal nature of the language and regional dialects complicate ASR. The AISHELL dataset provides 178 hours of Mandarin speech from 400 speakers, focusing on regional accents and the HKUST dataset contains over 200 hours of telephone conversation data, addressing informal speech [36] [37]. In Europe, low-resource languages like Welsh and Irish face similar challenges. The Mozilla CommonVoice project aims to support ASR development for these languages by providing large-scale, open-source speech data [38].

A. Dialectal Transcription in Swiss German

Swiss German, a group of Alemannic dialects spoken in the German-speaking regions of Switzerland, presents a notable example of these transcription challenges. Spoken by over five million people, Swiss German encompasses a variety of dialects, each with distinct phonetic, lexical, and grammatical characteristics [39]. However, the absence of a standardized writing system, combined with its diglossic nature, where Standard German predominates in written communication,complicates automatic transcription efforts [40].

Addressing these challenges requires substantial and diverse datasets, and the research community has made significant progress in developing resources for Swiss German ASR. Some key datasets include:

- STT4SG-350 is a Swiss German speech corpus comprising 350 hours of recordings, providing a balanced representation of all Swiss German dialect regions. The data was collected through a web application, where 316 speakers translated Standard German sentences into Swiss German. Demographic information, including dialect, age group, and gender, is also included in the dataset [41].
- SwissDial is a parallel corpus that includes approximately three hours of high-quality audio recordings for each of eight major Swiss German dialects, alongside corresponding Swiss German and Standard German transcripts. The dataset was developed by manually translating sentences sourced from newspapers and Wikipedia into the various dialects and recording them in a controlled studio environment [42].
- The Swiss Parliament Corpus (SPC) contains 293 hours of transcribed parliamentary speeches from the Swiss National Council and Council of States, primarily in the Bernese dialect. The corpus was created through forced sentence alignment, which pairs Swiss German speech with standard German text transcripts [43].

These datasets collectively aim to address the challenges of Swiss German transcription by providing essential resources for both training and evaluation purposes. However, significant challenges remain, particularly in achieving high accuracy across the diverse array of Swiss German dialects. Additionally, the scarcity of spontaneous conversational data adds complexity to the development of ASR systems that are capable of handling real-world speech contexts effectively.

IV. Comparative Analysis of Whisper and Wav2Vec 2.0

This section presents a comparative analysis of two state-of-the-art ASR models, Whisper and Wav2Vec 2.0, focusing on their application to Swiss German dialect transcription. Architectural differences, performance comparisons, and key research findings are examined to provide a comprehensive overview of the strengths and weaknesses of each model in this challenging domain.

A. Architecture

Whisper and Wav2Vec 2.0, while both state-of-the-art ASR models, diverge significantly in their architecture and training methodologies.

1) Whisper: OpenAI's Whisper is a transformer-based ASR model that utilizes an encoder-decoder architecture. It was trained on a massive, multilingual dataset consisting of 680,000 hours of supervised data, making it highly versatile across multiple languages and tasks. The training data includes a diverse range of languages, accents, and transcription types, as well as translation data, which enhances Whisper's ability to perform not only ASR but also translation tasks. This multilingual and multitask training approach gives Whisper a significant advantage when dealing with cross-lingual transfer, where knowledge gained from high-resource languages can be applied to low-resource ones. Whisper processes audio input by converting it into a log-Mel spectrogram, which captures both frequency and time information, providing a visual representation of sound. This spectrogram is then passed through the encoder, which consists of multiple transformer layers with self-attention mechanisms. The self-attention mechanism enables the model to capture temporal relationships between sounds, preserving context throughout the processing of the input. The encoder generates a sequence of contextualized embeddings, with each embedding representing a segment of audio and encoding information about its surrounding context. These embeddings are then passed to the decoder, which generates the textual output in an autoregressive manner. The decoder generates one token at a time, utilizing both the encoder's embeddings and previously generated tokens. The autoregressive process, facilitated by cross-attention between the decoder and encoder, ensures the model produces coherent transcriptions, even for lengthy or complex speech inputs. Whisper's flexibility is further enhanced by the use of special tokens that differentiate between tasks such as language identification, translation, and transcription.

2) Wav2Vec 2.0: In contrast, Meta AI's Wav2Vec 2.0 takes a self-supervised learning approach, which allows it to learn directly from raw audio without needing labeled transcriptions. This method makes Wav2Vec 2.0 more data-efficient, as it can be pre-trained on vast amounts of unlabeled audio data before being fine-tuned on a smaller set of labeled data for specific ASR tasks. This self-supervised pre-training is achieved through contrastive learning, where the model is trained to distinguish between correct and incorrect representations of speech features, helping it learn rich speech representations

even without direct supervision. The architecture of Wav2Vec 2.0 begins by processing raw audio through a convolutional neural network (CNN). The CNN is responsible for extracting low-level acoustic features from the raw waveform, such as spectral patterns, which are then quantized into discrete representations. These quantized features are fed into a transformer encoder, which captures long-range dependencies and contextual information within the audio sequence. The transformer architecture allows the model to build an understanding of how different sounds relate to each other over time, improving its ability to recognize speech accurately. For ASR tasks, Wav2Vec 2.0 requires fine-tuning on labeled data, typically using a Connectionist Temporal Classification (CTC) loss function, which enables the model to learn the alignment between speech signals and transcriptions. During the fine-tuning phase, Wav2Vec 2.0 is trained to optimize the CTC loss, ensuring that the output sequence of characters aligns with the corresponding input speech. Additionally, a diversity loss is incorporated into the training process to ensure that the model uses a broad set of codebook entries, improving its ability to generalize across different speech patterns.

B. Implications for ASR in Low-Resource Languages

These architectural and training differences have implications for the application of these models, particularly in scenarios with limited data.

1) Whisper: Whisper's ability to leverage a large multilingual and multitask dataset makes it highly effective in tasks that involve high-resource languages. Its pre-training on a diverse set of languages and tasks allows it to generalize well to a variety of ASR and translation tasks, even in the absence of extensive domain-specific data. This makes Whisper particularly advantageous for dialects where cross-lingual transfer learning can be applied, allowing knowledge from high-resource languages to assist in transcription tasks for low-resource dialects like Swiss German. However, Whisper's reliance on large, diverse datasets and substantial computational resources can pose limitations in resource-constrained environments, making it less practical in settings with limited access to computational power or data.

2) Wav2Vec 2.0: In contrast, Wav2Vec 2.0's self-supervised learning approach offers distinct advantages for situations where labeled data is limited or difficult to obtain. By training on vast quantities of unlabeled audio data, Wav2Vec 2.0 reduces the dependency on large amounts of transcribed speech, which is particularly valuable for low-resource languages and dialects. The model can be fine-tuned on smaller labeled datasets, making it more adaptable to ASR tasks in languages like Swiss German, where the availability of labeled data is scarce. This self-supervised method is especially beneficial for languages with limited linguistic resources, as it maximizes the utility of unlabeled audio to perform speech recognition tasks effectively.

C. Performance: The Swiss German Case Study

While ASR has made considerable progress in commercial applications, high error rates continue to pose a challenge, particularly in LVCSR [44]. Traditional evaluation metrics like Word Error Rate (WER) and BLEU score provide useful insights into ASR performance but are limited, especially for low-resource languages and spontaneous speech. WER evaluates the alignment between reference and predicted text at the word level, without considering semantic meaning, while BLEU, originally designed for machine translation, focuses on n-gram precision and fluency but falls short in capturing meaning preservation. Moreover, ASR errors, such as substitutions, deletions, and insertions, can adversely affect subsequent tasks like machine translation. Additional factors such as speaker variability, recording conditions, and language complexity further hinder ASR performance. To address these limitations, new evaluation methods that prioritize semantic accuracy, potentially leveraging large language models (LLMs), could offer a more comprehensive measure of ASR system effectiveness, ensuring that these systems are assessed not only on their transcription accuracy but also on their ability to preserve the meaning and intent of the spoken content.

Several studies have investigated the performance of state-of-the-art speech recognition models on Swiss German dialects, revealing valuable insights into the challenges and opportunities in this domain. Schraner et al. fine-tuned the XLS-R 1B Wav2Vec2 model on three Swiss German datasets: SDS-200, SPC, and SwissDial, demonstrating the effectiveness of this model for dialect transcription [39]. Plüss et al. explored the impact of data balance on performance by fine-tuning the same model on a balanced subset of the SwissDial dataset [41]. Their findings highlighted the importance of ensuring equal representation of different dialects within training data to improve overall accuracy. Dolev et al. evaluated Whisper Large-v3 in a zero-shot setting on STT4SG-350, SwissDial, and SPC datasets, showcasing its potential for cross-lingual transfer learning and generalization to unseen dialects [45].

The results in Table I highlight the importance of model architecture and training data on ASR performance for Swiss German dialects. Wav2Vec2 models consistently outperform Whisper Large-v3 across all datasets in both WER and BLEU scores. For instance, Schraner et al.'s fine-tuned Wav2Vec2 model achieved a WER of 15.23% and a BLEU score of 72.2, while Plüss et al.'s model reached a WER of 14.0% and a BLEU score of 74.7. In contrast, Whisper Large-v3 demonstrated higher error rates, with a WER of 23.0% and a BLEU score of 63.1 on STT4SG-350, indicating that

Table I
PERFORMANCE COMPARISON OF SPEECH RECOGNITION MODELS ON SWISS GERMAN DIALECT

Dataset	Model	WER	BLEU
STT4SG-350	Wav2Vec (Schraner et al.)	15.23%	72.2
	Wav2Vec (Plüss et al.)	14.0%	74.7
	Whisper Large-v3 (Dolov et al.)	23.0%	63.1
SPC	Wav2Vec (Plüss et al.)	23.7%	60.7
	Whisper Large-v3 (Dolov et al.)	29.5%	57.0
SwissDial	Whisper Large-v3 (Dolev et al.)	23.0%	61.0

fine-tuning on domain-specific data is crucial for improving accuracy.

While Whisper's performance is not as strong as fine-tuned Wav2Vec2, its zero-shot capabilities offer value in scenarios where retraining on dialect-specific data is not feasible. Whisper Large-v3 achieved a WER of 23.0% and a BLEU score of 61.0 on the SwissDial dataset, demonstrating its ability to generalize to unseen dialects.

Beyond architecture and training data, model size also significantly influences performance, as highlighted by the results presented in Table II, derived from Sicard et al. [46].

Whisper models exhibit improved performance as their number of parameters increases, with larger models consistently achieving better results. The improvements in both WER and BLEU score across the SPC and Swiss Dial datasets illustrate a consistent trend, irrespective of the dataset. For example, when comparing the extremes, Whisper Large (1.55B parameters) achieves a WER of 29.4% and a BLEU score of 56.2 on Swiss Dial, whereas Whisper Tiny (39M parameters) exhibits a substantially higher WER of 80.2% and a BLEU score of only 10.3. This difference underscores the advantage of larger models in capturing complex patterns and delivering more accurate transcriptions. Intermediate models further illustrate this trend. Whisper Base (140M parameters) achieves a WER of 66.8% and a BLEU score of 19.0, while Whisper Small (769M parameters) improves these metrics to a WER of 46.8% and a BLEU score of 36.4. Whisper Medium (1024M parameters) continues this progression, achieving a WER of 33.8% and a BLEU score of 50.8.

The choice between Whisper and Wav2Vec 2.0 depends on the availability of labeled data and computational resources. Wav2Vec 2.0, with its self-supervised pre-training, performs well with smaller labeled datasets, making it ideal for low-resource settings. It leverages large amounts of unlabeled data to learn representations, reducing the need for extensive labeled data. In contrast, Whisper requires a larger amount of labeled data due to its supervised learning approach, but it can achieve higher accuracy when sufficient data is available. Therefore, to improve Whisper's computational efficiency, Parameter-Efficient Fine-Tuning (PEFT) can be used, adjusting only a small subset of parameters, which reduces both computational and storage demands. Additionally, freezing some layers or adding specialized modules during fine-tuning can enhance Whisper's performance, particularly for low-resource languages [47] [48].

Table II
PERFORMANCE COMPARISON OF WHISPER MODELS

Dataset	Model	WER	BLEU
SPC	Whisper Tiny	96.3%	9.7
	Whisper Base	70.4%	20.3
	Whisper Small	50.9%	37.3
	Whisper Medium	37.4%	50.8
	Whisper Large	33.2%	55.6
Swiss Dial	Whisper Tiny	80.2%	10.3
	Whisper Base	66.8%	19.0
	Whisper Small	46.8%	36.4
	Whisper Medium	33.8%	50.8
	Whisper Large	29.4%	56.2

V. CONCLUSION AND FUTURE WORK

This paper examined the challenges and advancements in dialect transcription, focusing on the comparative performance of Whisper and Wav2Vec 2.0 for Swiss German. Our findings, drawn from existing research, reveal that fine-tuned Wav2Vec models generally outperform pre-trained Whisper in dialect-specific tasks, as evidenced by lower WER and higher BLEU scores. Despite its limitations in zero-shot scenarios, Whisper demonstrates potential for improvement, particularly when enhanced by fine-tuning techniques such as PEFT.

Future research should prioritize the development of more robust and accurate ASR systems that effectively address the challenges posed by data scarcity, lack of standardization, and grammatical divergence inherent in dialects. This includes exploring models like Whisper-turbo, which offer a compelling balance between size and performance, making them potentially well-suited for dialect transcription tasks. Moreover, the exploration of metrics beyond WER and BLEU, such as the use of LLMs to assess meaning preservation, is crucial for a comprehensive evaluation of ASR systems in challenging dialectal contexts. Another important area for future exploration is diarization, particularly in conjunction with the Whisper model. Enhancing Whisper's ability to handle speaker separation in multi-speaker scenarios will be crucial for practical applications in dialectal ASR, where multiple speakers often interact in varying linguistic styles.

In conclusion, significant progress has been made in dialect transcription, However, further research into model refinement, fine-tuning, and alternative evaluation techniques remains essential. Addressing the unique challenges of dialect transcription will enable the development of more accurate and reliable ASR systems capable of handling the wide variety of dialects present in spoken language.

REFERENCES

[1] H. Ibrahim and A. Varol, "A study on automatic speech recognition systems," in *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*, 2020, pp. 1–5.

[2] K. S and E. Chandra, "A review on automatic speech recognition architecture and approaches," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, pp. 393–404, 04 2016.

[3] D. Yu and L. Deng, *Automatic speech recognition*. Springer, 2016, vol. 1.

[4] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic speech recognition using advanced deep learning approaches: A survey," *Information Fusion*, p. 102422, 2024.

[5] A. Glasser, K. Kushalnagar, and R. Kushalnagar, "Deaf, hard of hearing, and hearing perspectives on using automatic speech recognition in conversation," in *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 2017, pp. 427–432.

[6] M. Vacher, A. Fleury, F. Portet, J.-F. Serignat, and N. Noury, "Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living," in *New Developments in Biomedical Engineering*, D. Campolo, Ed. In-Tech, 2010, pp. pp. 645 – 673, iSBN: 978-953-7619-57-2. [Online]. Available: https://hal.science/hal-00422576

[7] J. H. Hansen, X. Zhang, M. Akbacak, U. H. Yapanel, B. Pellom, W. Ward, and P. Angkititrakul, "Cu-move: Advanced in-vehicle speech systems for route navigation," *DSP for in-vehicle and mobile systems*, pp. 19–45, 2005.

[8] S. McCrocklin, "Asr-based dictation practice for second language pronunciation improvement," *Journal of Second Language Pronunciation*, vol. 5, no. 1, pp. 98–118, 2019.

[9] J. Jeon, S. Lee, and H. Choe, "Beyond chatgpt: A conceptual framework and systematic review of speech-recognition chatbots for language learning," *Computers & Education*, vol. 206, p. 104898, 2023.

[10] A. Bora and J. Vajpai, "Industrial applications of automatic speech recognition systems," *ournal of Engineering Research and Applications*, pp. 88–95, 03 2016.

[11] C.-C. Chiu, A. Tripathi, K. Chou, C. Co, N. Jaitly, D. Jaunzeikare, A. Kannan, P. Nguyen, H. Sak, A. Sankar *et al.*, "Speech recognition for medical conversations," *arXiv preprint arXiv:1711.07274*, 2017.

[12] J. G. Klann and P. Szolovits, "An intelligent listening framework for capturing encounter notes from a doctor-patient dialog," *BMC medical informatics and decision making*, vol. 9, pp. 1–10, 2009.

[13] G. S. Morrison, "Forensic voice comparison and the paradigm shift," *Science & Justice*, vol. 49, no. 4, pp. 298–308, 2009.

[14] D. Loakes, "Does automatic speech recognition (asr) have a role in the transcription of indistinct covert recordings for forensic purposes?" *Frontiers in Communication*, vol. 7, p. 803452, 2022.

[15] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[16] J.-C. Junqua and J.-P. Haton, *Robustness in automatic speech recognition: fundamentals and applications*. Springer Science & Business Media, 2012, vol. 341.

[17] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[18] Q. Liang, "Automatic speech recognition technology: History, applications and improvements," *Applied and Computational Engineering*, vol. 65, pp. 180–184, 05 2024.

[19] M. Gales and S. Young, "The application of hidden markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[20] A.-r. Mohamed, "Deep neural network acoustic models for asr." Ph.D. dissertation, University of Toronto Toronto, Canada, 2014.

[21] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr." in *Interspeech*, vol. 2012, 2012, pp. 22–25.

[22] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings 12*. Springer, 2015, pp. 91–99.

[23] A. Hannun, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[24] G. Zweig, C. Yu, J. Droppo, and A. Stolcke, "Advances in all-neural speech recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4805–4809.

[25] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[26] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[27] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for asr," *arXiv preprint arXiv:1904.11660*, 2019.

[28] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and lstm encoder decoder models for asr," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 8–15.

[29] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020. [Online]. Available: https://arxiv.org/abs/2005.08100

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[31] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[32] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[33] A. B. Wassink, C. Gansen, and I. Bartholomew, "Uneven success: automatic speech recognition and ethnicity-related dialects," *Speech Communication*, vol. 140, pp. 50–70, 2022.

[34] A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, and Y. Zhang, "The mgb-2 challenge: Arabic multi-dialect broadcast media recognition," 2019. [Online]. Available: https://arxiv.org/abs/1609.05625

[35] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra *et al.*, "Multilingual and code-switching asr challenges for low resource indian languages," *arXiv preprint arXiv:2104.00235*, 2021.

[36] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[37] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "Hkust/mts: A very large scale mandarin telephone speech corpus," in *International Symposium on Chinese Spoken Language Processing*. Springer, 2006, pp. 724–735.

[38] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[39] Y. Schraner, C. Scheller, M. Plüss, and M. Vogel, "Swiss german speech to text system evaluation," *arXiv preprint arXiv:2207.00412*, 2022.

[40] C. Paonessa, Y. Schraner, J. M. Deriu, M. Hürlimann, M. Vogel, and M. Cieliebak, "Dialect transfer for swiss german speech translation," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://openreview.net/forum?id=2X5RXTOsLU

[41] M. Plüss, J. Deriu, Y. Schraner, C. Paonessa, J. Hartmann, L. Schmidt, C. Scheller, M. Hürlimann, T. Samardžić, M. Vogel *et al.*, "Stt4sg-350: A speech corpus for all swiss german dialect regions," *arXiv preprint arXiv:2305.18855*, 2023.

[42] P. Dogan-Schönberger, J. Mäder, and T. Hofmann, "8swissdial: Parallel multidialectal corpus of spoken swiss german9, mar. 2021," *URL http://arxiv. org/abs/2103.11401*, 2021.

[43] M. Plüss, L. Neukom, and M. Vogel, "Germeval 2020 task 4: Low-resource speech-to-text." in *SwissText/KONVENS*, 2020.

[44] R. Errattahi, A. El Hannani, and H. Ouahmane, "Automatic speech recognition errors detection and correction: A review," *Procedia Computer Science*, vol. 128, pp. 32–37, 2018.

[45] E. L. Dolev, C. F. Lutz, and N. Aepli, "Does whisper understand swiss german? an automatic, qualitative, and human evaluation," *arXiv preprint arXiv:2404.19310*, 2024.

[46] C. Sicard, V. Gillioz, and K. Pyszkowski, "Spaiche: Extending state-of-the-art ASR models to Swiss German dialects," in *Proceedings of the 8th edition of the Swiss Text Analytics Conference*. Neuchatel, Switzerland: Association for Computational Linguistics, Jun. 2023, pp. 76–83. [Online]. Available: https://aclanthology.org/2023.swisstext-1.8/

[47] C. Kwok, S. Li, J. Qi, and E. Chng, "Low-resource language adaptation with ensemble of peft approaches," 2024.

[48] Y. Liu, X. Yang, and D. Qu, "Exploration of whisper fine-tuning strategies for low-resource asr," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 29, 2024.

Analysis of a Stochastic Predator-Prey Model for Detective-Criminal Dynamics

Viktor Najdovski
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University
Skopje, North Macedonia
viktornajdovski@students.finki.ukim.mk

Verica Bakeva
Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, North Macedonia
verica.bakeva@finki.ukim.mk

***Abstract*—This paper examines a stochastic density-dependent predator-prey model introduced by Cristina Gutiérrez and Carmen Minuesa. The main goal is analyzing the dynamics between detectives and criminals. We calibrate the model using violent crime reports and detective employment data from Washington (1985-2023) and perform a global variance-based sensitivity analysis. The results highlight the maximum survival probability of criminals as the key parameter driving criminal long-term growth rates and oscillation variability. Policies targeting this parameter would be most effective for reducing crime.**

***Index Terms*—Stochastic Processes, Predator-Prey, Detective-Criminal Dynamics, Sensitivity Analysis**

I. Introduction

Since their introduction by Lotka and Volterra, predator-prey models have demonstrated their utility in modeling many situations where one population acts as a predator and the other as its prey. However, most models are deterministic, based on ordinary differential equations.

We aim to model the population dynamics between detectives and criminals using a stochastic predator-prey model, where the detectives are considered as predators and criminals as their natural prey. Our goal is to derive policy recommendations through sensitivity analysis that effectively lower crime rates.

II. Probability Model

The probability model we will use in this paper was introduced by Cristina Gutiérrez and Carmen Minuesa in [1]. In the evolution of the process, we differentiate two phases:

- control phase, which models the interactions between the detectives and criminals
- reproduction phase, where every population recruits additional members

A predator-prey density-dependent branching process is a discrete time stochastic process $\{X, \tilde{X}\}_{n=0}^{+\infty}$ defined as:

$$(X_0, \tilde{X}_0) = (x_0, \tilde{x}_0)$$

$$(X_{n+1}, \tilde{X}_{n+1}) = \left(\sum_{i=0}^{\varphi_n(X_n, \tilde{\phi_n}(\tilde{X}_n))} \xi_i^{(n)}, \sum_{i=0}^{\tilde{\varphi}_n(X_n, \tilde{\phi_n}(\tilde{X}_n))} \tilde{\xi}_i^{(n)} \right)$$

for all $n \in \mathbb{N}_0$, where $(x_n, \tilde{x}_n)$ are non-negative integers. The random variables satisfy the following conditions:

i. The random variables of the family $\{\xi_i^{(n)}, \tilde{\xi}_i^{(n)}, \varphi_n(x, \tilde{x}), \tilde{\varphi}_n(x, \tilde{x}) : n, x, \tilde{x} \in \mathbb{N}_0, i \in \mathbb{N}\}$ are independent and non-negative integer valued.
ii. Within their respective families, $\{\xi_i^{(n)} : n \in \mathbb{N}_0, i \in \mathbb{N}\}$ and $\{\tilde{\xi}_i^{(n)} : n \in \mathbb{N}_0, i \in \mathbb{N}\}$ are independent and identically distributed with distributions $p = \{p_k\}_{n=0}^{+\infty}$, $\tilde{p} = \{\tilde{p}_k\}_{n=0}^{+\infty}$, mean values μ, $\tilde{\mu}$ and variance σ^2, $\tilde{\sigma}^2$ respectively.
iii. For $x, \tilde{x} \in \mathbb{N}$ and $n \in \mathbb{N}_0$, the variable $\varphi_n(x, \tilde{x})$ follows a binomial distribution with size x and probability of success $r\left(\frac{\tilde{x}}{x}\right)$, where $r : [0, \infty) \to (0, 1)$ is a continuous and strictly increasing function.
iv. For $x, \tilde{x} \in \mathbb{N}$ and $n \in \mathbb{N}_0$, the variable $\tilde{\varphi}_n(x, \tilde{x})$ follows a binomial distribution with size $\tilde{x}$ and probability of success $\tilde{r}\left(\frac{\tilde{x}}{x}\right)$, where $\tilde{r} : [0, \infty) \to (0, 1)$ is a continuous and strictly increasing function.

We introduce additional conditions for the functions $r(\cdot)$ and $\tilde{r}(\cdot)$. We assume the existence of some constants $0 < \rho_1 < \rho_2 < 1$, $0 < \tilde{\rho}_1 < \tilde{\rho}_2 < 1$, and $\gamma > 0$ satisfying:

v. $\lim_{x\to\infty} r(x) = \rho_2$ and $\lim_{x\to\infty} \tilde{r}(x) = \tilde{\rho}_2$
vi. $r(0) = \rho_1$ and $\tilde{r}(0) = \tilde{\rho}_1$
vii. The random variable $\varphi_n(x, 0)$ follows a binomial distribution with size x and probability of success ρ_1, and $\tilde{\varphi}_n(x, 0) = 0$ almost surely, for each $n, x \in \mathbb{N}_0$.
viii. The random variable $\tilde{\varphi}_n(0, \tilde{x})$ follows a binomial distribution with size $\tilde{x}$ and probability of success $\tilde{\rho}_2$, and $\varphi_n(0, \tilde{x}) = 0$ almost surely, for each $n, x \in \mathbb{N}_0$.
ix. $r(\gamma) = \frac{1}{\mu}$ and $\tilde{r}(\gamma) = \frac{1}{\tilde{\mu}}$

An example of such a function, which we will use, is provided in [1]:

$$r(x) = (\rho_2 - \rho_1)(1 - k^{-x}) + \rho_1, \text{ where } k = \left(\frac{\rho_2\mu - \rho_1\mu}{\rho_2\mu - 1} \right)^{1/\gamma}$$

$$\tilde{r}(x) = (\tilde{\rho}_2 - \tilde{\rho}_1)(1 - k^{-x}) + \tilde{\rho}_1, \text{ where } k = \left(\frac{\tilde{\rho}_2\tilde{\mu} - \tilde{\rho}_1\tilde{\mu}}{\tilde{\rho}_2\tilde{\mu} - 1} \right)^{1/\gamma}$$

Since there is not an infinite amount of resources in the world, we introduce a carrying capacity parameter $K > 0$ and define a random variable $\tilde{\phi}_n(\tilde{x})$, which follows a binomial distribution with size $\tilde{x}$ and probability of success $\tilde{s}(\tilde{x}, K)$. The function $\tilde{s}(\cdot, K) : [0, \infty) \to [0, 1]$ is continuous and strictly

decreasing depending on the parameter K. For computational efficiency we utilize the Beverton-Holt model given by the expression:

$$\tilde{s}(\tilde{x}, K) = \frac{K}{K + \tilde{x}}, \quad \text{for } K > 0$$

III. Model Comparison

We analyze monthly homicide and aggravated assault reports in Washington from 1985 to 2023. These violent crime reports represent the "criminal" population, while the yearly number of employed detectives over the same period serves as the "predator" population in our model.

Since we only have aggregated crime report counts, denoted by $\tilde{X}$, we must estimate the underlying offspring distribution $\tilde{\xi}_i$. To do this, we examine the growth ratios:

$$\frac{\tilde{X}_{t+1}}{\tilde{X}_t} \quad t = 0, 1, \ldots, n-2$$

We fit various candidate distributions to these ratios and evaluate them using Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). Table I summarizes the results:

TABLE I
BIC values for various distributions

Distribution	BIC	AIC	AICc
norm	-785.9234	-794.2161	-794.19
gamma	-784.1519	-796.5909	-796.53
lognorm	-784.0544	-796.4934	-796.44
weibull min	-781.8599	-794.2989	-794.24
t	-779.7771	-792.2161	-792.16
beta	-778.4998	-795.0852	-794.99
uniform	-454.4933	-462.7859	-462.00
expon	-259.5615	-267.8541	-267.00

Additionally, the Shapiro-Wilk and Kolmogorov-Smirnov tests both support normality:

$$\text{Shapiro-Wilk Test: Statistic} = 0.99, \ \text{p-value} = 0.10$$
$$\text{Kolmogorov-Smirnov Test: Statistic} = 0.04, \ \text{p-value} = 0.42$$

Based on the BIC/AIC metrics and normality tests, we select normal distribution as the best fit. The BIC and AIC values for lognormal and gamma are very close (difference of $\sim 1-2$), so normal distribution is only marginally better. Heavy-tail distributions, such as lognormal or gamma, are almost as plausible.

The estimated parameters for the normal distribution are $\mu = 1.006$ and $\sigma = 0.1$. However, the estimated mean understates the actual growth rate due to external factors such as resource limitations and law enforcement efforts. To adjust this, we incorporate losses from carrying capacity (l_K) and predation (l_p), scaling the mean as follows:

$$\mu_{\text{adjusted}} = \mu \frac{1}{l_p \, l_K} = 1.006 \frac{1}{0.7 \cdot 0.84} \approx 1.71$$

Analyzing clearance data for Washington cases, we find that, on average, approximately 57% of cases are solved. To account for additional deterrence effects, we assume that 14% of criminals cease criminal activity in the next generation without being captured, influenced by law enforcement presence. This assumption, while not empirically derived, is introduced to reasonably align the model dynamics. Combining these factors, we set $l_p = 0.7$. For l_K, we start by fixing the carrying capacity to $K = 6000$ and compute the survival rate as:

$$l_K = s(\bar{\tilde{X}}, K) = \frac{K}{K + \frac{1}{n}\sum_{i=0}^{n-1} \tilde{X}_i} \approx 0.84$$

Since we only have yearly data about the number of employed detectives in Washington, we cannot perform statistical tests to conclude the offspring distribution. Therefore, we simplify the model even further by assuming the detectives offspring distribution follows a normal distribution, similar to that of the criminals, with mean 1.8 and standard deviation 0.2, based on the fact that the observed ratio of violent crime reports to detectives is around 0.712.

Based on historical data from Washington, we set the model parameters as follows:

- density dependence parameter (γ) is set to 0.712, reflecting the above mentioned observed ratio of violent crime reports to detectives in Washington.
- Criminals' Maximum Survival Probability ($\tilde{\rho}_2$): Recognizing that criminals rarely face complete suppression, we set $\tilde{\rho}_2 = 0.8$, reflecting a high survival probability.
- Detectives' Maximum Survival Probability (ρ_2): Considering factors such as workload, potential retirements, injuries, and budget constraints, not all detectives remain active indefinitely. To account for these factors, we set $\rho_2 = 0.8$.
- Minimum Survival Probabilities (ρ_1 and $\tilde{\rho}_1$): At low densities, it is reasonable to assume that some detectives continue their duties, and some criminal activity persists, as crime rates have historically never dropped to zero. Therefore, we set both ρ_1 and $\tilde{\rho}_1$ to 0.2.

With these parameters, we simulate the model and examine the detective-criminal population dynamics under varying values of K, illustrated in Figure 1.

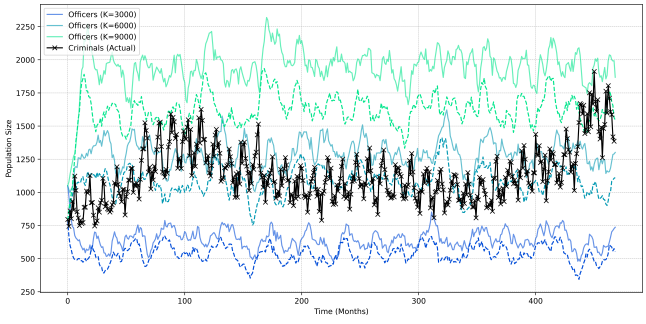


Fig. 1. Detective-Criminal simulation under varying K

The simulation shows a clear density-dependent relationship between detectives and criminals, in which the criminal population oscillates around the mean of the observed criminal re-

ports. An increase in criminal activity leads to a corresponding rise in the detective population, highlighting greater demand for law enforcement. Conversely, as crime rates fall, the need for law enforcement diminishes, resulting in a reduced number of detectives.

To justify the complexity of our stochastic model, we compare it to the simpler deterministic Lotka-Volterra ODEs and observe that the deterministic approach does not fully capture the stochastic nature of real-world crime data. Instead, it produces regular, predictable oscillations without reflecting the inherent randomness of the system.

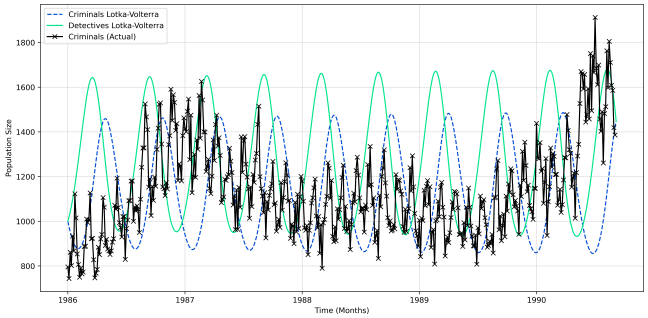


Fig. 2. Lotka-Volterra simulation

IV. Sensitivity Analysis

To understand which parameters most strongly influence criminal population dynamics, we conduct a global variance-based sensitivity analysis. Having fixed the distributions of criminals and detectives and the carrying capacity, we examine how varying each parameter affects key outputs of the model. This helps identify which policy interventions—represented by parameter shifts—are likely to have the greatest impact on crime reduction.

Threshold analysis indicates that parameter values of $\tilde{\rho}_2$ and ρ_2 below 0.6 lead to system extinction, a scenario inconsistent with real-world crime dynamics. Therefore, we constrain our global sensitivity analysis for these parameters to the range $[0.6, 1.0]$ to focus on parameter influence under non-extinction conditions. The other parameters $\tilde{\rho}_1$ and ρ_1 are also extinction-safe for their respective ranges.

We apply Sobol sensitivity analysis, which decomposes the variance of the model output into contributions from each input parameter and their interactions. The parameters analyzed are:

$$\begin{aligned} \gamma &: \text{ ranged in } [0.5, 3.0] \\ \rho_1, \tilde{\rho}_1 &: \text{ ranged in } [0.0, 0.4] \\ \rho_2, \tilde{\rho}_2 &: \text{ ranged in } [0.6, 1.0] \end{aligned}$$

Each parameter set is sampled using Saltelli's method with $N = 8192$ base samples, ensuring high accuracy in estimating both first and total-order Sobol indices. For each sampled set, we simulate 468 generations, discard the initial transient period (first 234 generations), and analyze the model's stable behavior. Each simulation is repeated five times to account for stochastic effects.

A. Criminal Growth Rates

The growth rate reflects how rapidly crime escalates or diminishes over time. By analyzing sensitivity with respect to growth, we are pinpointing which parameters control long-term trends. If a parameter has a strong influence on growth rates, changing it accordingly can lower crime rates. The results are presented below in Table II. The maximum survival

TABLE II
Sobol Sensitivity Indices and Confidence Intervals (Growth Rates)

Parameter	S_1	**95% CI for** S_1	S_T	**95% CI for** S_T
γ	0.036	[0.023, 0.049]	0.173	[0.160, 0.185]
ρ_1	0.094	[0.078, 0.111]	0.284	[0.269, 0.298]
ρ_2	0.009	[0.002, 0.017]	0.065	[0.059, 0.072]
$\tilde{\rho}_1$	0.065	[0.049, 0.080]	0.249	[0.234, 0.264]
$\tilde{\rho}_2$	0.536	[0.514, 0.558]	0.733	[0.714, 0.753]

probability of criminals, $\tilde{\rho}_2$, is the primary driver of variance in criminal population growth. It accounts for approximately 53% of the variance by itself and about 73% when accounting for interactions with other parameters. This highlights that policies aimed at reducing the maximum survival probability of criminals—whether through improved detection technologies or improved policing strategies—are the most effective in halting long-term criminal growth.

The parameters ρ_1 and $\tilde{\rho}_1$, representing the minimum survival rates, also exhibit a moderate influence, with total Sobol indices around 0.25-0.28. In contrast, γ and ρ_2 contribute minimally.

Lowering $\tilde{\rho}_2$ not only reduces the criminal population but also indirectly leads to a decline in the number of detectives, as reduced criminal activity decreases the demand for law enforcement. The impact of reducing $\tilde{\rho}_2$ from 0.8 to 0.65 is shown in the simulation below, with all other parameters configured identically to the simulation in Figure 1.

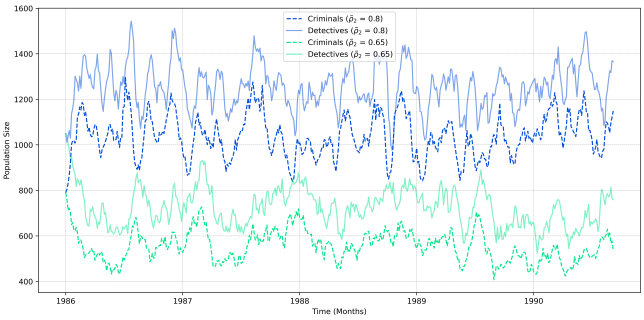


Fig. 3. Effect of lowering $\tilde{\rho}_2$ on criminal and detective populations

After running 1000 simulations, discarding the transient state, we get the following results:

$$\begin{aligned} &\text{Average crime level for } \tilde{\rho}_2 = 0.8 : 1043.99 \\ &\text{Average crime level for } \tilde{\rho}_2 = 0.65 : 552.88 \\ &\text{Percentage reduction in average crime} : 47.04\% \end{aligned}$$

B. Criminal Oscillation Variability

Oscillation amplitude reflects instability and unpredictability of crime levels. Large oscillations mean unpredictable spikes, which are harder for detectives to manage. Stable crime levels lead to easier planning and resource allocation for detectives. The measure of amplitude we use is:

$$\max\{\tilde{X}_t : t \in [n/2, n-1]\} - \min\{\tilde{X}_t : t \in [n/2, n-1]\}$$

TABLE III
SOBOL SENSITIVITY INDICES AND CONFIDENCE INTERVALS (OSCILLATION AMPLITUDE)

Parameter	S_1	**95% CI for** S_1	S_T	**95% CI for** S_T
γ	0.017	[-0.002, 0.036]	0.355	[0.313, 0.397]
ρ_1	0.030	[0.009, 0.051]	0.467	[0.422, 0.512]
ρ_2	-0.002	[-0.021, 0.017]	0.299	[0.265, 0.333]
$\tilde{\rho}_1$	0.026	[0.004, 0.048]	0.417	[0.382, 0.452]
$\tilde{\rho}_2$	0.356	[0.328, 0.384]	0.733	[0.693, 0.773]

We observe that the maximum survival probability of criminals, $\tilde{\rho}_2$, dominates again. Fixing its value reduces over 70% of the variance in oscillation amplitude. The significant difference between first-order and total-order indices for all parameters indicates strong interaction effects. By themselves, the rest of the parameters have limited impact; however, in combination they considerably influence oscillation behavior.

We also note that small negative S_1 values indicate noise; they are treated as zero within statistical error. The results from lowering $\tilde{\rho}_2$ from 0.8 to 0.65 are visualized below:

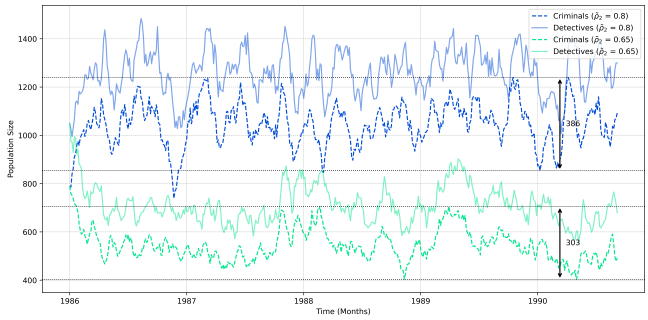


Fig. 4. Effect of lowering $\tilde{\rho}_2$ on criminal amplitude

After running 1000 simulations, discarding the transient state, we get the following results:

$$\text{Average amplitude for } \tilde{\rho}_2 = 0.8 : 500.23$$

$$\text{Average amplitude for } \tilde{\rho}_2 = 0.65 : 394.15$$

$$\text{Percentage reduction in amplitude} : 21.21\%$$

V. LIMITATIONS AND FUTURE WORK

Modeling the detective-criminal dynamics is a difficult task. The analysis of the model offers valuable insights; however, limitations should be acknowledged:

- The model assumes that every new criminal originates from an already existing criminal. The mean of the criminal offspring distribution might seem high to account for this factor. Introducing an additional random variable to represent new criminals not derived from existing ones and adding it to the criminals' offspring distribution might address this limitation.
- We treat all of Washington state as one homogeneous population. Crime and detective resources are not evenly distributed—Seattle's dynamics differ from rural areas. Future work might segment the data.
- The model currently attributes all changes to predator-prey interactions and random noise. If a major policy like Washington's "Three Strikes" law in the 90s impacted crime, the model treats it as an internal fluctuation. Introducing a random variable that models such changes or incorporating time-varying parameters could be two potential solutions.
- The model has fixed parameters over the 38-year period; this is unrealistic in the ever-evolving world we live in. Incorporating time-varying parameters would address this issue.
- While the sensitivity analysis identifies key parameters influencing criminal population dynamics, translating these findings into concrete policy recommendations is difficult, requiring extensive domain knowledge.

VI. CONCLUSION

In this paper, we used actual crime data from Washington from 1985 to 2023 to analyze the dynamic relationship between detectives and criminals using a stochastic predator-prey model. We demonstrated that stochastic models capture the inherent randomness in crime rates more realistically than traditional deterministic approaches.

Through sensitivity analysis we identified that lowering the maximum criminal survival rate has the greatest impact in suppressing long-term crime growth through policies such as better crime investigation technology and better police strategies, followed by increasing the minimal detective survival rate.

While our model simplifies certain real-world complexities, it nonetheless provides valuable insights into how law enforcement resources can be strategically allocated.

REFERENCES

[1] C. Gutiérrez and C. Minuesa, "Predator–prey density-dependent branching processes," *Stochastic Models*, vol. 38, no. 4, pp. 537–570, 2022.
[2] U.S. Bureau of Labor Statistics, "Occupational Employment and Wage Statistics Tables." [Online]. Available: https://www.bls.gov/oes/tables.htm. Accessed: Mar. 18, 2025.
[3] Federal Bureau of Investigation, "Crime Data Explorer: Crime Trends – Washington State." [Online]. Available: https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/crime-trend. Accessed: Mar. 18, 2025.
[4] A. J. Lotka, *Elements of Physical Biology*. Baltimore, MD, USA: Williams and Wilkins, 1925.
[5] V. Volterra, "Fluctuations in the abundance of a species considered mathematically", *Nature*, 1926.
[6] A. Saltelli, K. Chan, and E. M. Scott, *Sensitivity Analysis*. New York, NY, USA: Wiley, 2000.

Web-Based Platform for Enhancing Coordination in Technology Transfer

Marija Karajanova
Faculty of Comp. Sci. and Eng.
Ss. Cyril and Methodius University
Skopje, North Macedonia
marija.karajanova@students.finki.ukim.mk

Dimitar Kitanovski
Faculty of Comp. Sci. and Eng.
Ss. Cyril and Methodius University
Skopje, North Macedonia
dimitar.kitanovski@finki.ukim.mk

Zorica Karapancheva
Faculty of Comp. Sci. and Eng.
Ss. Cyril and Methodius University
Skopje, North Macedonia
zorica.karapancheva@finki.ukim.mk

Aleksandar Stojmenski
Faculty of Comp. Sci. and Eng.
Ss. Cyril and Methodius University
Skopje, North Macedonia
aleksandar.stojmenski@finki.ukim.mk

Petre Lameski
Faculty of Comp. Sci. and Eng.
Ss. Cyril and Methodius University
Skopje, North Macedonia
petre.lameski@finki.ukim.mk

Ivan Chorbev
Faculty of Comp. Sci. and Eng.
Ss. Cyril and Methodius University
Skopje, North Macedonia
ivan.chorbev@finki.ukim.mk

***Abstract*—Nowadays, many traditional processes need digital automatization. Overload of paperwork, sending e-mails and inability to track work activities are a few of the problems occurring in professional environments. Working in challenging circumstances can influence productivity which is of great importance when it comes to exchanging technologies. A necessity for an automized platform is more than evident. For that purpose, we constructed this paper which aims to present a software solution as a web-based platform for better coordination of technology transfer. Essentially, the main idea behind the proposed solution is to solve the workflow centralization issue and improve collaboration between expert professionals and external industry entities.**

***Index Terms*—technology transfer, digitalization, knowledge-sharing, collaboration, client, consultant**

I. Introduction

In solving problems, the idea has always played a key role. However, the idea itself in its initial form does not necessarily lead to an invention. Without converting these ideas into real solutions, there is a high possibility that they will be left on paper and forgotten over time. Many global efforts have been made to build the stages of materializing innovations into a unified process, leading to the concept called technology transfer. Naumovska defines the term *technology transfer* as "the process of transferring scientific discoveries, technologies, skills, knowledge, production methods and products from one institution or country to another for the purpose of further development and commercialization" [1]. In modern times, the question remains about its implementation which, if done successfully, can lead to an increase in institutional capabilities.

Since this concept is not completely new, it is constantly being improved to respond to new requirements and solve developing problems. Digitalization has been a current issue in the modernization of the process recently. Many digital platforms offer assistance; however, dividing the work across different platforms is not the most optimal solution. Therefore, we propose a solution that is initially developed following the stakeholders' requirements of INNOFEIT, which is the first fully operational Digital Innovation Hub (DIH) in North Macedonia. By giving businesses access to technical expertise and networking opportunities, INNOFEIT promotes innovation through digital technology. This allows businesses to test before investing and speeds up the commercialization of new solutions.

The system discussed in this paper offers several key functionalities such as project management, application management, activity tracking, discussion and feedback mechanism, document sharing and time tracking. In addition to process automation, it provides explicit and clear insight into the merits of both parties, the time invested and the overall contribution, ensuring transparency throughout the entire period. The platform is designed with the flexibility to adapt to similar challenges in related domains, providing an interchangeable structure.

II. Literature Overview

Over the past years, many theoretical frameworks have been discussed for supporting a successful technology transfer process in innovation and entrepreneurship studies. A foundational model remains the Triple Helix model introduced in 1997 by Henry Etzkowitz and Loet Leidesdorf. The focus is mainly on the collaboration between academic institutions, industries, and government, as to how theoretical expertise will contribute to practical problems. A study conducted for this model describes the university-industry-government relationship and entails that academia can take an entrepreneurial role in the innovation and economic growth of the industry with governmental support of any kind [2]. Moreover, it introduces the importance of centralization of university-industry linkage with technology transfer offices (TTOs) and collaborative research centers for strengthening the technology transfer process. The role of TTOs is significantly important in university-industry interaction as it institutions an organizational focus on

formal mechanisms of commercialization such as patenting, licensing and entrepreneurship [3].

More recent studies have concentrated on the digital transformation of the technology transfer process in the university's TTOs, its importance, and its respective outcomes. Doering and Timinger answer two key questions - why is there a need for digitalization of knowledge transfer as part of the universities' third mission and how can this process be presented in a structured framework [4]. They explain that a digital transfer process will automize the workflow leading to reducing administrative and organizational issues of research and project transfer, as well as improvement of internal and external services. The universities not only that can safeguard the transfer, but also will be able to handle needs from external partners for easier and faster knowledge transfer, especially with SMEs who require quick solutions. A recent paper also explains that using digital tools enhances communication, productivity, and collaboration which makes the process more efficient [5], however, concerns such as security risks and digital imbalances still need to be addressed.

Ferro-Serrano et al. explore data collected from Horizon Europe and the Digital Economy and Society index in [6]. They concluded that countries with intermediary levels of digitalization are more effective in technology transfer as they constantly work on improving their position in the network. This indicates that a digital tool should support the process rather than completely replace the traditional practices of knowledge-sharing.

III. Methodology

The methodology consists of a system overview, an explanation of designing the system architecture, as well as additional integration and database schema.

A. System Overview

The Technology Transfer System [or some other name we decide to use] is a web-based platform facilitating the collaboration between expert professionals (consultants) and external industry entities (clients) in the project workflow. The core functionality of the system consists of the ability for clients to propose projects with detailed descriptions and track their progress of execution by the consultants within a deadline. As the system serves three main actors - client, consultant, and administrator - it needs to be explained through different user role perspectives.

- **Clients** such as external organizations, companies, institutions, etc. can define titled projects, categorize them by expertise, and set their deadlines and descriptions explaining the respective goals, requirements, and scope. After the client makes the project available on the platform, he waits for applicants to express interest in participating. He can either approve or reject applications in order to construct the expert team. The client has control over the project data integrity as he can edit its details and timeline, or even delete it if necessary. However, he does not directly contribute to the workflow and development of the project. Instead, the system allows him to monitor its progress and stay informed through regular updates.
- **Consultants** are expert professionals, academia staff, or research specialists whose domain-specific knowledge and experience, clients can effectively use as contributions to their projects. The system presents a consultant with a list of projects matching his expertise to which he can apply. The consultant is given an outlook of his applications from where he can withdraw pending requests or resubmit rejected ones. Once accepted, consultants gain access to the project's inside information and are responsible for defining the activities they will operate. As the project can become complex and disorganized, the creation of activities is necessary. The consultant specifies an estimated period for the activity's execution allowing clients to have clear insight into the real workflow of the project and be informed about the actual progress of completed work. With this approach the system enables a structured breakdown of work, ensuring assigned responsibilities and well-optimized prosecution.
- **Administrators** have insight into the overall state of the system and the data within it. They have full control over the workflow as they can approve or reject consultants' applications, and adjust projects' or activities' deadlines, progress, and other related information. Administrators can create projects and assign them to a specific client and a team of consultants. They can create activities as well for a specific project assigning them to a specific consultant who is part of the project's execution. Even though they cannot create or delete users from the system, they can however delete client's projects and remove consultants from a project.

One of the main goals of the system is to provide ongoing transparency and visibility without compromising user trust. Thus, all users involved in the project besides administrators can leave comments about the activities with the ability to upload attachments within the comment. Users can edit and delete their comments and shared documents. They can also download any attachments uploaded by their coworkers. Administrators are deprived of this information as otherwise it would directly violate privacy and intellectual property rights. Not only does this feature facilitate collaboration between peers involved in the same project, but also acts as a feedback mechanism for clients who cannot directly impact the project but are still able to control its progress. Additionally, the system calculates the time each consultant spends working on the project based on the timespan assigned for each activity. This information is updated in real-time. Compared to earlier approaches that depended on document reviews, self-reporting, and emails to approximately measure job engagement, this is a major advance as the long manual procedure is now automated by the system in seconds.

B. Database Design

After retrieving the requirements for the real-world objects' relationships, a conceptual data model diagram is developed

using techniques such as the entity-relationship model shown in Fig. 1. ER model is often used to successfully address a wide variety of design concerns in a straightforward form [7].

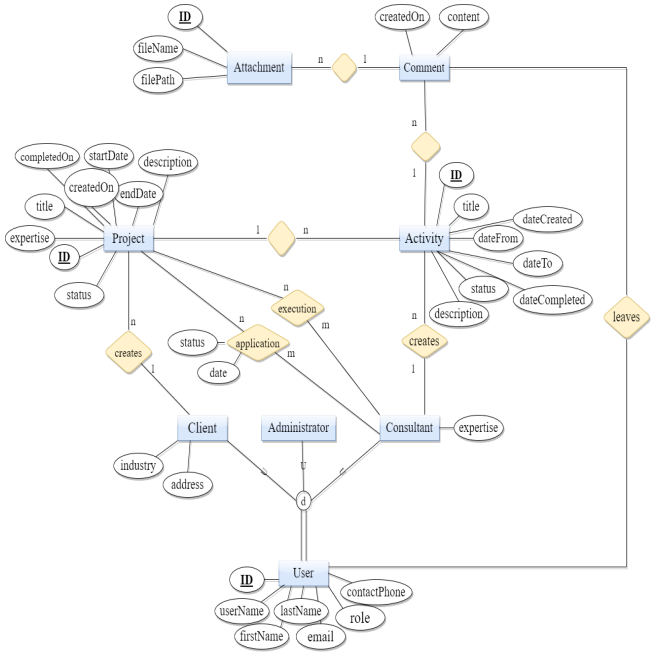


Fig. 1. ER diagram

In this system, records of clients, consultants, projects, activities, attachments, and comments are represented as entities. Each entity in the database is distinguished by a unique identifier as its primary key preventing duplication of tuples in the tables. Clients, consultants, and administrators are users who share common attributes such as name, surname, email, phone number, and password. Clients have an additional attribute categorizing the background industry while consultants are categorized by expertise. The project entity has a central role in the system with main attributes including title, description, start date, end date, and status. The activity entity which serves as the project's main operational unit, has characteristics corresponding to the project entity. Comments are associated with activities and users, containing attributes such as body and creation date, while attachments are linked to the comments, with attributes such as file name and file path ensuring traceability of the file resource.

The creation of a project by a client is represented as a one-to-many relationship where the client can create multiple projects and the project has a single creator. From the system overview, it is evident that the consultant-project interaction consists of two key stages: the application stage where a consultant applies for a project, and the execution stage where a consultant is working on a project. As shown in Fig. 1 there are two separate relationships: application and execution. The application relationship includes two attributes, status and date of applying.

The diagram in Fig. 1 illustrates the Activity's relationships with the Consultant and Project as one-to-many, showcasing that a consultant creates multiple activities and each activity is linked to a specific project. However, for more accurate representation Fig. 2 depicts a revised overview where the execution relationship is converted to an entity and the activity entity links to it. The main idea is to imply that an activity is created for a particular project by a consultant who is already involved in that project. A rule states that a consultant cannot create an activity for a project he is not working on, therefore this relationship model guarantees that the business logic is maintained.

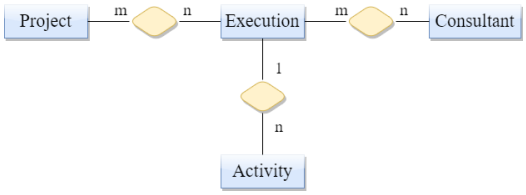


Fig. 2. consultant-project-activity revised overview

C. System Architecture

In order to achieve a scalable yet organized web-based system, the concept follows a client-server architecture [8] integrating MVC pattern [9] and onion architecture [10].

a) Client-Server Architecture: Client-Server Architecture is a structure for distributed applications that divides workloads between clients, who request services, and servers, who supply those services. When the client computer submits an HTTP(S) request to the web server via the Internet, the server acknowledges the request and transmits the desired files for the website. The browser on the client side then renders the files through DOM interpreter, CSS interpreter, and JS Engine, so the client has access to the website. Our concern is regarding the server-side application, as according to client-server architecture it handles the main business logic, operations, and user requests.

b) MVC Pattern: The Model-View-Controller (MVC) design pattern is used to decompose the server-side application into three components: Model, View, and Controller, each serving a different purpose. The sole responsibility of the Model focuses on data management, which evolves around business rules processing and data manipulation according to the Controller's requests. The actual data needs to be displayed as organized content assembled intuitively for the user to understand and additional data entered by the user needs to flow into the application. The View presents the data consistent with the Model through the user interface and brings the user input for further propagation to the Controller. Any logic implemented in the View, besides page structure, relates to presenting relevant content to the user. These components need to be coordinated appropriately which defines the functionality of the Controller. It handles the application logic when the user requests a page or enters input for processing by updating

the Model and the View accordingly to reflect the occurred modifications.

c) Onion Architecture: Onion architecture is an architectural pattern very similar to layered architecture therefore emphasizing the separation of concerns. The main concept is that it enforces inward coupling instead of tight coupling; all coupling is towards the center. At the center, there is the Domain Model and around it there are other layers whose number can vary. In our system, other layers consist of the Repository layer, Business layer, and Representation layer. Domain, Repository, and Business layers represent the application core. Additionally, two other principles are covered with the usage of Onion architecture complementing the benefit of loose coupling: Dependency Inversion and externalization of infrastructure which in our case represents the database. In the aspect of Dependency Inversion, every layer will provide its services through interfaces implemented during runtime using the concept of Dependency Injection, while in the aspect of externalization of the infrastructure, the database is an external storage system that communicates with the application core through an interface. The integration of MVC and Onion is simple as we equate the application core with the Model from the MVC pattern, while Controller and View components are placed as part of the Representation layer.

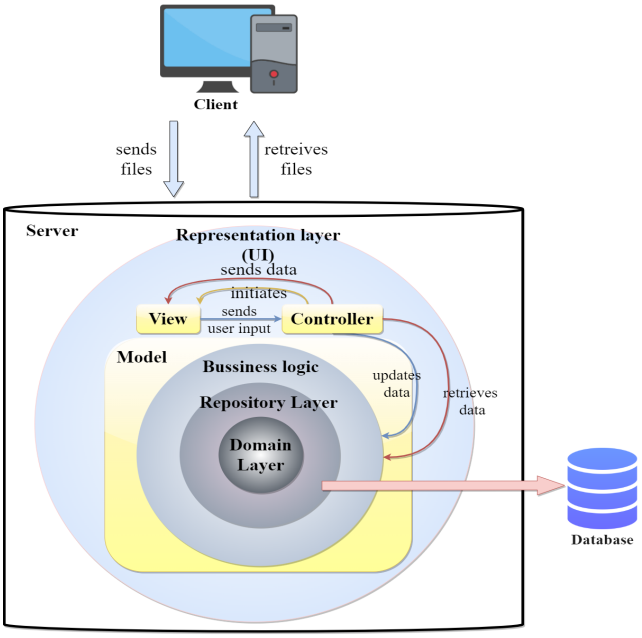


Fig. 3. System architecture

The system in Fig. 3 illustrates a visual representation of how our system coherently integrates previously explained strategies. Additionally it is separated into two modules. The distinction of two separate applications is mainly because of the system's nature. Administrators have different privileges and needs than other users in the system. The admin application simply follows the MVC architecture, which retrieves and modifies the necessary data through a RESTfull controller implemented in the main application.

D. API Communication

The admin application communicates with the main application through a controller that follows REST architecture [11], [12] shown in Fig. 4. If the traditional controllers serve views to clients, a RESTful controller enables the admin application to act as a client requesting resources rather than UI files. The system's universal data transfer format is JSON chosen for the simplicity of parsing to internal data objects. Both applications use data transfer objects converted in JSON which are sent using HTTPS protocol. Each application mapps the transferred object to appropriate domain entity and further uses it. The main application solely manages authorization and authentication, therefore when an administrator user logs in to the system, the main application redirects him to the admin application.

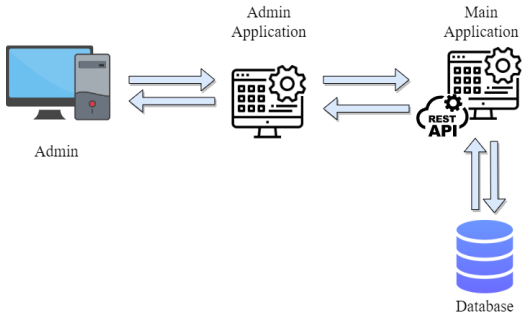


Fig. 4. API communication

IV. Discussion

The key idea of our work is to introduce a maintainable, scalable, and testable software solution while addressing challenges such as coupling and separation of concerns. The implementation is done using Microsoft technologies, including Microsoft SQL Server for managing the database, ASP.NET MVC Framework as a base technology, and ASP.NET Web API Framework for REST services.

Firstly, the choice to use an SQL database in this project comes from a variety of advantages that align with the needs of the system. As Hernandez explains in his book "Database Design for Mere Mortals: A Hands-On Guide to Relational Database Design", a relational database provides data and relationship integrity which guarantees data accuracy and consistency, as well as easy data retrieval in various ways either from a single table or multiple related tables [13]. Additionally, relational databases provide the independence of logical and physical data from database applications, which means that changes made to the logical design or physical implementation do not negatively impact the system. Although systems built upon relational databases ran very slowly in the past, the rapid advancement of hardware and software

engineering has significantly fastened processing and database transactions, enabling the full potential of relational databases.

Secondly, a key advantage of using the MVC pattern is the separation of concerns. Each component can be developed independently by different individuals or teams without breaking the concept. The division of responsibilities not only helps in achieving parallelism in real-time development but also diminishes the complexity so it is easier to code, debug, and test. Onion architecture contributes to the issue of separation of concerns, but also coupling between components. It introduces loose coupling as every class can use a different object from another class without being dependent on it. MVC and Onion have mostly similar aims as architectural patterns, so implementing them together coherently provides manageable scalability. Additionally, REST architecture for Web APIs gives a standardized way for modules to communicate without being tightly coupled in the meantime.

Furthermore, the separation of modules is an initial step toward migrating to a microservice architecture. As business logic evolves, a common approach in software development is converting to microservices for better management of individual components where each can follow integration of various architectures. If microservices are put into practice carefully and correctly, the system changes quickly and safely at the same time. This means more agility, and more agility means better outcomes [14].

V. Conclusion

Contemporary events suggest that the development of centralized platforms like the one proposed in this paper is essential in the transition from innovation to real-world implemented invention. The objective of this paper stands to introduce centralization and automatization of the core procedures within the technology transfer process. Key challenges that were taken into perspective with the highest priority were workflow fragmentation between different digital tools and lack of transparency. The system addresses these challenges by incorporating several functionalities starting from project management, task management, document sharing, communication, and feedback mechanisms to efficient time tracking and record keeping. An additional objective we also attempted to address was better collaboration between the experts and the external stakeholders providing a unified platform of this kind where each has insight not only of individual contribution but team share as well. We can propose that the platform can be reused in facilitating other processes sharing similar features according to their necessities and demands. Integration of several architectural frameworks based on MVC, Onion, and REST principles for modular separation along with SQL relational database for data management was designed to ensure a scalable and maintainable solution open for future modifications in terms of the requirements' nature. It was also taken into account the scalability of the system introducing a step forward to microservice architecture which is a modern practice providing room for further growth.

As the digitalization of manual processes becomes an inevitable part of moving forward, the technology transfer process itself is not an exception. Under the condition of continuously adapting to the evolving needs of the technology exchange process, [technology transfer system or some other name] can promise to make a significant impact in the acceleration of innovation and knowledge exchange not only in the university-industry partnership but in various domains as well.

References

[1] L. Naumovska, *Handbook for Technology Transfer*, 1st ed., 03 2022.

[2] A. T. Hailu, "The role of university–industry linkages in promoting technology transfer: implementation of triple helix model relations," *Journal of Innovation and Entrepreneurship*, vol. 13, pp. 1–17, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269033863

[3] M. Perkmann, V. Tartari, M. McKelvey, E. Autio, A. Broström, P. D'Este, R. Fini, A. Geuna, R. Grimaldi, A. E. Hughes, S. Krabel, M. Kitson, P. Llerena, F. Lissoni, A. J. Salter, and M. Sobrero, "Academic engagement and commercialisation: A review of the literature on university-industry relations," *O&M: Structures & Processes in Organizations eJournal*, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:155029804

[4] C. Döring and H. Timinger, "Industry-oriented digital transformation in universities to facilitate knowledge transfer," in *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227129912

[5] R. M. Sarala, S. Y. Tarba, N. Zahoor, H. Khan, S. C. L. Cooper, and A. Arslan, "The impact of digitalization and virtualization on technology transfer in strategic collaborative partnerships," *The Journal of Technology Transfer*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274249181

[6] M. Ferrer-Serrano, L. Fuentelsaz, and M. Gil-Lamata, "More digitalization does not always imply more technology transfer: an analysis within the horizon europe strategy," *The Journal of Technology Transfer*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:273206845

[7] T. J. Teorey, S. Buxton, L. Fryman, R. H. Gting, T. A. Halpin, J. L. Harrington, W. H. Inmon, S. Lightstone, J. Melton, T. Morgan, T. Nadeau, B. O'Neil, E. J. O'Neil, P. E. O'Neil, and M. Schneider, "Database design: Know it all," 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:32462923

[8] S. A. Hamid, R. Alaudeen, and D. A. Khamees, "What is client-server system: Architecture, issues and challenge of client -server system (review)," 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:249884777

[9] A. Majeed and I. Rauf, "Mvc architecture: A detailed insight to themodern web applications development," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:69241323

[10] J. Palermo, "The onion architecture: Part 1," jul 2008. [Online]. Available: https://jeffreypalermo.com/blog/the-onion-architecture-part-1/

[11] J. Yasmin, Y. Tian, and J. Yang, "A first look at the deprecation of restful apis: An empirical study," *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 151–161, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221376912

[12] M. Biehl, *API Architecture: The Big Picture for Building APIs*, ser. API University Series - Volume 2. API University, 2015.

[13] M. J. Hernandez, "Database design for mere mortals: A hands-on guide to relational database design (3rd edition)," 1996. [Online]. Available: https://api.semanticscholar.org/CorpusID:57035350

[14] M. Kratzke, B. H. K., and other authors, *Microservices Up and Running: A Step-by-Step Guide to Building a Microservices Architecture*. O'Reilly Media, 2020. [Online]. Available: https://www.oreilly.com/library/view/microservices-up-and/9781492075448/ch01.html

Modelling the Dynamics of Heterogenous Populations

Jana Hristova

Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, North Macedonia
jana.hristova@students.finki.ukim.mk

Verica Bakeva

Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, North Macedonia
verica.bakeva@finki.ukim.mk

***Abstract*—This paper examines a stochastic population model introduced by Di Crescenzo et al., which describes species competition for environmental sites through colonization, death, and replacement. The model includes a hierarchical structure, where species dominance influences site occupation. Theoretical properties are analyzed, including occupancy probabilities and stationary distributions, with simulations implemented to explore species coexistence dynamics. Additionally, real-world applications and limitations of the model are assessed, highlighting its relevance and constraints in ecological modeling.**

***Index Terms*—stochastic population models, species competition, birth-death processes, stochastic simulation**

I. Introduction

In this paper we examine a stochastic process that can be used to model heterogenous populations. In addition to a detailed theoretical explanation of the model, we also presnet a simulation of the model and explore its possible applications. Finally, an analysis of the model's limitations will be conducted, highlighting potential areas for improvement.

II. General Model

The model examined in this paper, proposed by Di Crescenzo, Giorno, Nobile, and Ricciardi, is a modified birth-death process designed to describe interactions in multi-species populations [1]. It incorporates three key factors—colonization (birth), death, and replacement—to model competition among species for habitat occupancy.

This section focuses on the general model, while the next section explores a specific case of the model that incorporates species hierarchy.

A. Rules of Evolution

The population inhabits an environment divided into many sites, each occupied by at most one individual from s different species $1, 2, \ldots, s$, where $s \geq 2$. The dynamics at each site follow these rules:

i. Suppose a given site is empty at time t. Then, in the time interval $(t, t+\Delta t]$, the site is either occupied by an individual of species j $(j = 1, 2, \ldots, s)$ with probability $\lambda_j \Delta t + o(\Delta t)$ or remains empty with probability $1 - \sum_{j=1}^{s} \lambda_j \Delta t + o(\Delta t)$.

ii. Suppose a given site is occupied by an individual of species i at time t. Then, in the time interval $(t, t+\Delta t]$, the individual leaves the site and it is unoccupied with probability $\mu_i \Delta t + o(\Delta t)$, the individual is replaced by an individual of another species j $(j \neq i)$ with probability $\nu_{ij}\Delta t + o(\Delta t)$, or the individual continues to occupy the site with probability $1 - \mu_i \Delta t - \sum_{\substack{j=1 \\ j\neq i}}^{s} \nu_{ij}\Delta t + o(\Delta t)$.

B. Occupancy of Environmental Sites

Let $\{X(t);\, t \geq 0\}$ be a stochastic process representing the occupancy of a given environmental site, where $X(t) = 0$ if the site is unoccupied and $X(t) = j$ if the site is occupied by an individual of species j. Then, from the two rules of evolution in the previous section, the following can be obtained:

$$P\{X(t+\Delta t) = j \mid X(t) = 0\} = \begin{cases} 1 - \sum_{i=1}^{s} \lambda_i \Delta t + o(\Delta t), & \text{if } j = 0 \\ \lambda_j \Delta t + o(\Delta t), & \text{if } j = 1, 2, \ldots, s \end{cases} \tag{1}$$

$$P\{X(t+\Delta t) = j \mid X(t) = i\} = \begin{cases} \mu_i \Delta t + o(\Delta t), & \text{if } j = 0,\ i = 1, 2, \ldots, s \\ \nu_{ij}\Delta t + o(\Delta t), & \text{if } i, j = 1, 2, \ldots, s,\ i \neq j \\ 1 - \mu_i \Delta t - \sum_{j=1, j\neq i}^{s} \nu_{ij}\Delta t + o(\Delta t), & \text{if } i, j = 1, 2, \ldots, s,\ i = j \end{cases} \tag{2}$$

Additionally, we assume that events occurring in disjoint time intervals are independent.

Now, we consider an environment divided into N independent sites. Let $\{Y(t) = [Y_1(t), Y_2(t), \ldots, Y_s(t)];\, t \geq 0\}$ be the s-dimensional stochastic process that describes the number of sites occupied by each species. Consequently, $\{Y_r(t);\, t \geq 0\}$ is the one-dimensional stochastic process that represents how many of the N sites are occupied by individuals of species r $(r = 1, 2, \ldots, s)$. If $X_k(t) \quad (k = 1, 2, \ldots, N)$ denotes the occupancy of site k at time t, the following is obtained:

$$Y_r(t) = \sum_{k=1}^{N} \delta[X_k(t), r] \quad (r = 1, 2, \ldots, s), \tag{3}$$

where

$$\delta[X_k(t), r] = \begin{cases} 0 & \text{if } X_k(t) \neq r, \\ 1 & \text{if } X_k(t) = r. \end{cases} \quad (4)$$

If the N sites in the environment are empty at time $t = 0$, the probability that, at time t, N sites are occupied by $n_1, n_2, \ldots, n_s$ individuals of species $1, 2, \ldots, s$, respectively, is given by:

$$p(n_1, n_2, \ldots, n_s; t) = \binom{N}{n_1}\binom{N-n_1}{n_2}\cdots\binom{N-\sum_{i=1}^{s-1} n_i}{n_s} \times [q_1(t)]^{n_1}[q_2(t)]^{n_2}\cdots[q_s(t)]^{n_s} \times [q_0(t)]^{N-\sum_{i=1}^{s} n_i}. \quad (5)$$

$$p^{(r)}(n; t) = \binom{N}{n} [q_r(t)]^n [1 - q_r(t)]^{N-n} \quad (6)$$

Equation (5) corresponds to a multinomial distribution, while equation (6) corresponds to a binomial distribution which is also the marginal distribution.

C. Model Characteristics

From (5) and (6), the following four equations are derived and described accordingly.

$$E[Y_r(t) \mid Y_r(0) = 0] = Nq_r(t) \quad (r = 1, 2, \ldots, s) \quad (7)$$

$$\text{Var}[Y_r(t) \mid Y_r(0) = 0] = Nq_r(t)\,[1 - q_r(t)] \quad (r = 1, 2, \ldots, s) \quad (8)$$

Equations (7) and (8) represent the expected value and variance of the number of sites occupied by species r.

$$E[Y_i(t)Y_j(t) \mid Y_i(0) = 0, Y_j(0) = 0] = N(N-1)q_i(t)q_j(t) \quad (i, j = 1, 2, \ldots, s; i \neq j). \quad (9)$$

$$\text{Cov}[Y_i(t), Y_j(t) \mid Y_i(0) = 0, Y_j(0) = 0] = -Nq_i(t)q_j(t) \quad (i, j = 1, 2, \ldots, s; i \neq j). \quad (10)$$

Equation (9) represents the expected number of distinct site pairs, where one site is occupied by species i and the other by species j. On the other hand, the covariance (Equation 10), which is negative in this case, indicates competition between species for resources (sites), as an increase in one species' occupancy leads to a decrease in the other's.

III. Model with Species Hierarchy

In this section, an upgrade to the general model is considered, where each species is associated with a position in a hierarchy. In this scenario, the death rates are constant, and the rates of replacement of individuals depend only on the species j that colonizes, with replacements allowed only if the colonizing species has a higher rank (a lower index) than the species that already occupies the environmental site. With this, it can be stated that individuals of species with higher ranks have a higher priority for colonization, and thus can be considered privileged. The assumptions described above for this version of the general model can be mathematically expressed as follows:

$$\lambda_i > 0, \quad \mu_i = \mu > 0 \quad (i = 1, 2, \ldots, s) \quad (11)$$

$$\nu_{ij} = \begin{cases} \nu_j \geq 0, & \text{if } i > j \\ 0, & \text{if } i \leq j \end{cases} \quad (i, j = 1, 2, \ldots, s) \quad (12)$$

From this, it can be concluded that $\nu_{is} = 0$ for all $i = 1, 2, \ldots, s$, since species s has the lowest rank and therefore cannot be a colonizer.

In order to be able to express the stationary probabilities, the following cumulative rates are introduced:

$$L_i := \sum_{k=1}^{i-1} \lambda_k \quad (i = 2, 3, \ldots, s) \quad (13)$$

$$L \equiv L_{s+1} := \sum_{k=1}^{s} \lambda_k \quad (14)$$

$$V_i := \sum_{k=1}^{i-1} \nu_k \quad (i = 2, 3, \ldots, s) \quad (15)$$

Equation (13) represents the total rate of colonization by all individuals coming before individual i. This describes the influence of the previous species in the colonization process. Equation (14) is the total rate of colonization of all species. Finally, equation (15) represents the total rate of replacement resulting from species before species i, that is, how much the species before i contribute to the replacement of an individual in a given environmental site.

A. Stationary Distribution

In order to determine the stationary behavior of the population within the model, the stationary probabilities are represented as follows:

$$q_i = \lim_{t \to \infty} q_i(t) \quad (i = 0, 1, \ldots, s) \quad (16)$$

As a result, the following stationary distribution is obtained:

$$q_0 = \frac{\mu}{\mu + L} \quad (17)$$

$$q_i = \frac{\mu}{\mu + L} \frac{\nu_i(L - L_i) + \lambda_i(\mu + V_i)}{(\mu + V_i)(\mu + V_{i+1})} \quad (18)$$

$$q_s = \frac{\mu}{\mu + L} \frac{\lambda_s}{\mu + V_s} \quad (19)$$

From equations (17), (18), and (19), it can be observed that an increase in the colonization rate λ_j decreases the probability that a specific environmental site remains empty, while increasing the probability that it will be occupied by species j. As a result, the probability of occupation by less privileged species decreases. When dealing with more privileged species, the effect of increasing λ_j depends on whether $\frac{\nu_i}{\mu + V_i}$ is greater

or less than $\frac{\lambda_i}{\mu+L_i}$: If the former is greater, the probability of occupation by species i increases; otherwise, it decreases. An increase in the death rate increases the probability that a specific site remains unoccupied and decreases the probability of occupation by the most privileged species. On the other hand, an increase in the replacement rate ν_j does not affect the probability that the site remains unoccupied, reduces the probability of occupation by less privileged species, and does not cause any change when dealing with more privileged species.

IV. Model Simulation

To facilitate a better understanding of the model, this chapter will present a general simulation of the model using Python. The simulation will be performed on two systems: one with three species and another with five species, each with different colonization and replacement rates. The species are ranked hierarchically, which, as mentioned earlier, means that a species with a higher rank (lower index) can replace a species with a lower rank (higher index), but not vice versa.

The simulation utilizes the Gillespie algorithm, a stochastic simulation method, to track the number of locations occupied by each species over time [2]. The algorithm calculates the colonization, death, and replacement rates and accordingly selects and executes events based on these rates. The simulation runs for a fixed time period (in this case, $time_{max} = 1000$), and the state of the system (the number of locations occupied by each species and the number of empty locations) is recorded in each event. The results of both simulations are visualized in graphs that show the number of locations occupied by each species over time. These graphs illustrate how the system evolves toward a stable state.

A. Simulation with Five Species

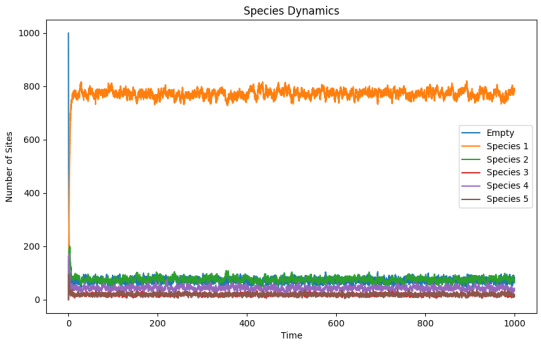


Fig. 1. Dynamics of scenario with five species

In the scenario shown in Figure 1, we consider a population consisting of five species of individuals and an environment divided into 1000 sites. Each species has a death rate of $\mu = 0.1$, while the corresponding colonization and replacement rates are as follows: $\lambda_1 = 0.2, \lambda_2 = 0.3, \lambda_3 = 0.1, \lambda_4 = 0.4, \lambda_5 = 0.3, \nu_1 = 0.4, \nu_2 = 0.2, \nu_3 = 0.1, \nu_4 = 0.3, \nu_5 = 0.0$. Within the simulation, the theoretical and simulated stationary probabilities are calculated and presented in Table 1. Additionally, the mean absolute error is computed as $MAE = 0.0025$, indicating that the simulation adequately represents the stochastic process, allowing conclusions to be drawn from the graph. As expected, species 1, which has the highest privilege, occupies the largest proportion of sites in the environment, while all other species have occupied significantly fewer sites. In fact, all remaining species have occupied a similar or lower proportion of sites compared to the proportion of empty sites. The graph also shows that this system reaches its stationary distribution in a very short time.

Occupancy	Theoretical probabilities	Simulated probabilities
Unoccupied (q_0)	0.0714	0.0721
Species 1 (q_1)	0.7714	0.7720
Species 2 (q_2)	0.0661	0.0751
Species 3 (q_3)	0.0139	0.0184
Species 4 (q_4)	0.0430	0.0429
Species 5 (q_5)	0.0195	0.0196

TABLE I
Comparison of theoretical and simulated probabilities

B. Simulation with Three Species

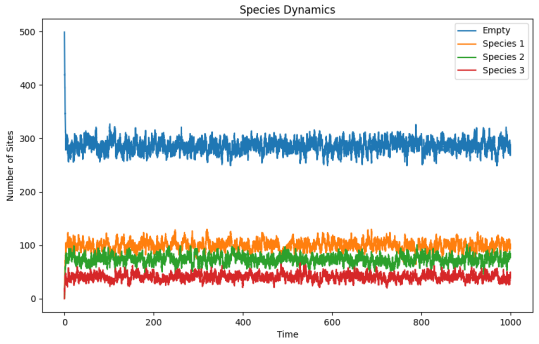


Fig. 2. Dynamics of scenario with three species

Occupancy	Theoretical probabilities	Simulated probabilities
Unoccupied (q_0)	0.5714	0.5725
Species 1 (q_1)	0.2000	0.1973
Species 2 (q_2)	0.1469	0.1477
Species 3 (q_3)	0.0816	0.0825

TABLE II
Comparison of theoretical and simulated probabilities

In the scenario shown in Figure 2, we consider a population consisting of three species of individuals and an environment divided into 500 sites. Each species has a death rate of $\mu = 0.4$, while the corresponding colonization and replacement rates are as follows: $\lambda_1 = 0.1, \lambda_2 = 0.1, \lambda_3 = 0.1, \nu_1 = 0.1, \nu_2 = 0.2, \nu_3 = 0.0$. The corresponding theoretical and simulated stationary probabilities are presented in Table 2,

with a mean absolute error $MAE = 0.0014$. Compared to the previous scenario with five species, here the species have a higher death rate, so most sites remain unoccupied, as can also be observed in Figure 2, where the blue color represents the number of empty sites. Since all three species have the same colonization rate and species 1 has a lower replacement rate, the species occupy a similar number of sites, unlike in the previous case, where species 1 had the highest replacement rate and the highest priority, resulting in a significantly larger number of occupied sites compared to the other species.

In this scenario, as expected based on the hierarchy, species 1 still occupies the largest number of sites, but this time, due to the equal colonization rate, the difference in the number of occupied sites is significantly smaller. As in the previous example, the system quickly reaches its stationary distribution.

V. Applications

i. Managing Invasive Species
The model can predict how invasive species replace native ones in an ecosystem. Species with higher colonization rates (λ_i) act as invaders, while hierarchical replacement highlights how dominant species outcompete others. This helps in assessing biodiversity loss and guiding management strategies.

ii. Metapopulation Dynamics
The model can be applied to study the dynamics of species spread and extinction in fragmented habitats, helping to understand the resilience of metapopulations. Metapopulations consist of small, spatially separated populations of the same species [3]. This model can be used to predict the settlement dynamics of such populations over time, taking into account factors such as colonization and local extinction. This could aid in designing conservation networks for metapopulations.

iii. Urban Growth and Land Use
The model can be utilized to simulate urban expansion, where different land uses (residential, commercial, green spaces, industrial) act as competing species. Colonization (λ_i) can represent the speed of development, while replacement (ν_i) can reflect land conversion, which aids urban planning.

iv. Gentrification and Social Mobility Another application of this model can be found in the realm of social mobility, where it can be used to capture gentrification dynamics. In this scenario, the species would be different socioeconomic groups, where higher-income groups colonize neighborhoods, replacing lower-income inhabitants. This type of use of the model would help analyze housing trends and guide policy decisions.

VI. Limitations

i. Uniform Death Rates
The model assumes equal death rates (μ_i) for all species, oversimplifying the real-world dynamics where the mortality of species varies due to seasons, diseases, food availability and other ecological factors.

ii. Ignoring Environmental Factors
Key environmental influences such as temperature, precipitation, and habitat destruction are not explicitly included in the model, although the model is ecological in its core. These overlooked factors can significantly affect survival, reproduction, and competition, limiting the applicability of the model in complex ecosystems.

iii. No Evolutionary Dynamics
The model does not consider genetic adaptation or species co-evolution. In reality, species evolve in response to environmental pressures, altering their competitive abilities, colonization rates, and interactions over time.

VII. Conclusion

In this paper, a stochastic model for the dynamics of heterogeneous populations was analyzed, focusing on species interactions and the effects of hierarchical structure. The results showed that competitive interactions and species ranking significantly influence colonization processes and ecosystem stability. The model provides a quantitative analysis of these processes, with potential applications in ecology and natural resource management. Future research could extend the model by incorporating spatial effects and external environmental factors.

References

[1] Di Crescenzo, A., Giorno, V., Nobile, A. G., and Ricciardi, L. M. "Stochastic population models with interacting species," *Journal of Theoretical Biology*, vol. 210, no. 1, pp. 111–126, 2001.

[2] "Stochastic Simulation," *BE150 - Principles of Biological Circuit Design*, California Institute of Technology, 2020. Accessed 4 Feb. 2025. http://be150.caltech.edu/2020/content/lessons/12_stochastic_simulation.html#Coding-up-the-Gillespie-simulation

[3] Encyclopædia Britannica, "Metapopulation," Encyclopædia Britannica, Inc. Available: https://www.britannica.com/science/metapopulation. [Accessed: 20-Mar-2025].

Exploration into LLMs capabilities for question-answering over tabular data

Teodora Mladenovska, Martina Toshevska, Sonja Gievska
Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering
Skopje, North Macedonia mladenovska.teodora@students.finki.ukim.mk,
{martina.toshevska, sonja.gievska}@finki.ukim.mk

***Abstract*—Answering questions over tabular data often involves using structured query languages or specialized tools, which adds complexity for individuals without expertise in data querying or technical skills. In this paper, we present different methods for answering questions over tabular data from SemEval-2025 Task 8, leveraging the capabilities of large language models (LLMs) and agent tools such as Langchain CSV Agent and Langgraph. While other approaches exist, such as using SQL agents or building a custom agent for interacting with database data, we limit our experiments to the two selected methods due to constraints in computational resources and cost. We compare the performance of both open-source(Mistral-7B-Instruct-v0.2, DeepSeek-R1-Distill-Qwen-1.5B) and proprietary model(Claude 3.5 Sonnet), evaluating their results to determine which approaches give the best results for this specific dataset. The results show that Claude 3.5 Sonnet with the CSV agent achieves the highest accuracy of 90.84%, while open-source models perform with lower accuracy. These findings highlight the strengths and limitations of various approaches for answering questions in tabular data, providing insight into the effectiveness of LLMs and agent tools for structured data interpretation.**

***Keywords*—agents, langgraph, proprietary models, open source models, tabular data**

I. Introduction

Structured (tabular) data is one of the most common data formats people use in various industries for day-to-day tasks. With the rapid growth of technology, the amount of data that is generated and collected has increased significantly [1]. This data can be organized in different ways, but one of the most common ways is that it is often represented in a tabular format, making it easier to store and process. The tabular format provides the advantage of easy analysis, and it can be exported from databases, spreadsheets, or other data management systems, enabling businesses and researchers to extract meaningful insights about a given subject [2].

However, specialized skills are required in order to analyze such data. A data analyst, scientist, or someone proficient in query languages or with experience in using a given tool, would be needed to extract relevant information, interpret trends, and generate insights. This requirement creates barriers for individuals in various fields, such as doctors and nurses in healthcare, educators in academia, financial advisors in banking, policymakers in government, and supply chain managers in logistics, who may lack technical expertise but still need to derive insights from the provided data.

With the emergence of Large Language Models (LLMs), we can now close this gap by enabling users to analyze structured data using natural language. Instead of writing complex queries or performing manual data exploration, individuals can ask questions about the provided data in plain language and receive fast, accurate, and meaningful results. This feature allows data now to be accessible by a single individual, without requiring deep technical knowledge [2].

II. Related Work

The authors in [3] demonstrate that Knowledge Graphs can enhance LLM-based question-answering systems for enterprise insurance SQL databases. The paper demonstrates that Knowledge Graph increases the system's accuracy from 16% to 54%. Additionally, the proposed approach detects and fixes incorrect queries, boosting accuracy to 72.55%.

Using LLM prompting, the conversion of natural language questions into Python functions is demonstrated in [4]. With this approach, the task is broken into smaller steps, clearly showing its reasoning process and including a self-correction mechanism to address execution errors.

On the other hand, [5] explores the ability of large language models (LLMs) to process Electronic Health Records (EHRs) for patient data extraction and retrieval. Using the MIMICSQL dataset, the research evaluates the impact of prompt structure, instruction, and context on the performance of Llama2 and Meditron. Results indicate that optimal feature selection and serialization can enhance performance by up to 26.79%, while in-context learning improves data extraction by 5.95%. The study provides guidelines for designing LLM-based models to support health-related search tasks.

[6] compares the performance of a traditional relational database and a financial knowledge graph with retrieval-augmented generation (RAG), focusing on stock prices and financial data from the SET50 index. The results show that the graph database retrieves results faster, while the queries generation process using LLMs is with the similar accuracy for both databases.

Furthermore, [7] introduces AutoTQA, an autonomous tabular question answering framework that uses multi-agent large language models (LLMs) to handle queries across multiple tables from various systems. The system includes five agents and employs scheduling algorithms, along with a low-code tool

called LinguFlow, achieving excellent performance on several datasets.

III. Materials and Methods

This study leverages state-of-the-art AI frameworks and large language models (LLMs) to process structured tabular data from CSV files. As we already mentioned, traditional tabular data analysis requires technical knowledge of query languages (such as SQL) or data processing libraries (such as Pandas in Python). However, the development of AI and agent-based reasoning systems enables smooth interaction with structured data through plain text. Using LangChain CSV Agents and LangGraph custom agents, we explore the potentials of LLM-driven tabular data analysis, incorporating both open-source and proprietary models [8], [9].

A. Tools

LangChain is an open-source framework that makes it easier to build applications with large language models (LLMs). It offers a consistent interface for creating chains and integrates with various tools [10].

1) Langchain CSV agents: One of the key components in Langchain are agents. Agents are systems designed to handle high-level tasks by using large language models (LLMs) as reasoning engines. These agents analyze the given task, make decisions about the necessary actions, and then execute those actions autonomously based on the input [11]–[13].

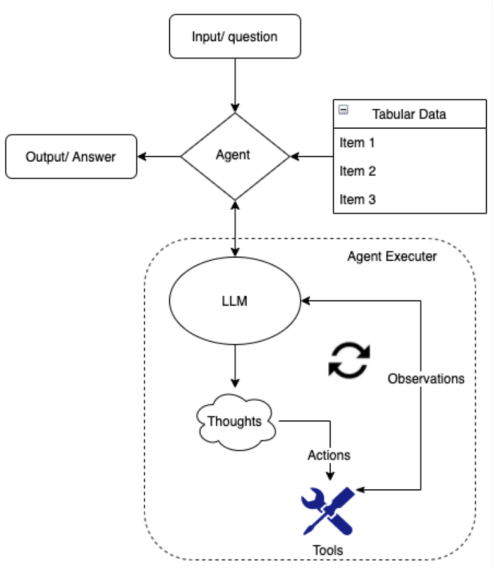


Fig. 1: Langchain CSV agents

The main components of Langchain CSV agent are:

- **Thought:** Before taking any action, the agent analyzes the question based on the provided information and creates a strategy to retrieve the relevant information. This step involves:
  - Understanding the question
  - Identifying necessary information (e.g., filtering, aggregation)
  - Ensuring that the approach is logical and aligns with the dataset's schema
- **Action:** Once the thought process is finished and the execution plan is created, the agent selects the appropriate tool (in this case, Python REPL tool), to answer the user's question.
- **Action Input:** The agent in this step translates the natural language question into executable Python code that interacts with the dataframe (df). It assumes that the action task is performed based on the thought process in the previous step and is getting processed by the selected tool. Since the agent is using the Python REPL tool, the input is expected to be valid Python commands. The correctness of this step is crucial because any syntax errors or incorrect logic can result in execution failures.
- **Observation (Optional Looping):** It contains the response from the action. In some cases, a single execution may not be efficient to answer the question, so the agent will iterate through the following steps:
  - Reassess the problem based on the observation.
  - Adjust the action or refine the input.
  - Execute another command.

  This iterative cycle continues until enough information has been gathered. The agent ensures that each new step brings it closer to a complete response.
- **Final Thought:** The agent determines that it has all the necessary information to answer the question and generates a final response.
- **Final Answer:** The clear, human-readable response is created.

Here is an example of how one execution chain looks like:

`> Entering new AgentExecutor chain...`
Let me help find the book with the lowest number of copies left that still has some copies in stock.
Thought: I need to look at the `'Copies Left'` column for books that are `'In Stock'` and find the minimum value.
Action: `python_repl_ast`
Action Input: `df[df['Stock Status'] == 'In Stock']['Copies Left'].min()`
Let me modify my approach to find the specific book with the lowest non-zero number of copies left.
Action: `python_repl_ast`
Action Input: `df[(df['Stock Status'] == 'In Stock') & (df['Copies Left'] > 0)].sort_values('Copies Left').head(1)[['Book Title', 'Copies Left']]`
I now know the final answer.
Final Answer: "The Startup Wife" is the book closest to running out of copies with just 1 copy left in stock.
`> Finished chain.`

2) Langgraph: LangGraph, an extension of LangChain, is specifically built to enhance the capabilities of agents by offering greater control and customization. It enables developers

to create agents that can be customized to fit specific needs, offering a more controlled way to tackle complex problems. With LangGraph, users can create agents that not only perform actions but also follow precise instructions, adapt to dynamic environments, and integrate seamlessly with various tools and external data sources. The LangGraph is composed of nodes, edges, and conditional edges. Nodes are units that perform specific tasks, edges are connections defining the flow between nodes, and conditional edges guide the flow based on conditions [14], [15].

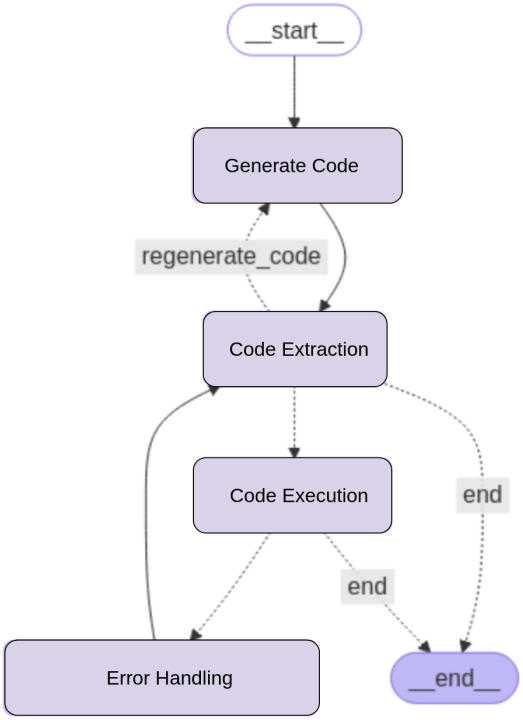


Fig. 2: Custom Langgraph CSV agent

The main components of our custom CSV agent are the following nodes:

- **Code Generation:** The first component of our agent is the Code Generation Node, which takes a user-provided query along with a table schema as input. Using these inputs, it generates Python code designed to answer the query. This node serves as the foundation of the agent, translating natural language questions into executable code.
- **Code Extraction:** The Code Extraction Node follows the generation step. The Python code generated during the Code Generation Node is usually provided within the following syntax: "'`python"'. It extracts the Python code from the output produced by the previous node, ensuring that only the necessary code is passed to the next node.
- **Code Execution:** The extracted Python code is passed to the Code Execution Node, where it is executed using the built-in `exec()` function.
- **Error Handling:** The final component is the Error Handling Node, which comes into play if any error occurs during the execution of the Python code. This node captures the error along with the relevant code and uses the error message to identify potential fixes. The error-handling mechanism attempts to correct the code and then sends it back to the Code Extraction Node for reprocessing.

The flow between the **Code Generation** and **Code Extraction Nodes** is defined by a standard edge. A conditional edge links the **Code Extraction** and **Code Execution Nodes**: if the code is successfully extracted, it is passed for execution. However, if extraction fails, the **Code Generation Node** is re-triggered to produce a new code snippet. The **Code Execution Node** also has a conditional path: if the code runs successfully, the process ends; if not, it proceeds to the **Error Handling Node**. The **Error Handling Node** receives the error and the code, attempts to fix the issue, and then sends the corrected code back to the **Code Extraction Node** through another edge for reprocessing. To prevent potential infinite loops, a limit is set on how many times the Code Extraction Node and Code Execution Node can be retried. If this limit is reached, the process is terminated to ensure the graph avoids endless iterations.

B. Models

During this research the following models were used:

1) Claude 3.5 Sonnet: demonstrates significant advancements in coding and tool use, achieving state-of-the-art performance on industry benchmarks such as SWE-bench Verified (49.0%) and TAU-bench (69.2% in retail, 46.0% in airline tasks). Early evaluations by GitLab and Cognition highlight its enhanced reasoning and efficiency in multi-step software development and autonomous AI workflows, making it a strong candidate for complex problem-solving and automation [16].

2) Mistral-7B-Instruct-v0.2: is an improved version of Mistral-7B, a powerful open-weight language model designed for instruction-following tasks. Built on a dense transformer architecture with 7 billion parameters, it offers strong performance in natural language understanding, reasoning, and code generation while maintaining efficiency [17].

3) DeepSeek-R1-Distill-Qwen-1.5B: is a small, open-weight, instruction-tuned language model developed by DeepSeek-AI, designed as a distilled version of larger models. It balances efficiency and performance, making it well-suited for reasoning, natural language processing, and general-purpose AI tasks [18].

IV. Results and Discussion

The dataset used in this research is part of the Semeval 2025-Task 8 competition [19]. It includes 15 tables and 522 questions about those tables. Each question is also linked to the name of the dataset it came from. The results of using different tools and models are shown in the table below. The accuracy score, which is used as the metric, represents the proportion of correct predictions made by the model out of the total number of predictions and is calculated using a script from the competition's repository [20].

Based on the Table I, we can conclude that the best results from our experiments come from using the Claude model with the CSV agent. The results of the CSV agent combined with the open-source Mistral model are much lower, clearly showing

Method	Model	Score
Langchain CSV agent	Claude 3.5 Sonnet	90.84
Langchain CSV agent	Mistral-7B-Instruct-v0.2	43.72
Llangraph	DeepSeek-R1-Distill-Qwen-1.5B	20.88
Llangraph	Claude 3.5 Sonnet	84.87

TABLE I: Comparison of different methods and models

that open-source models perform worse for our task due to their limited reasoning and function-calling capabilities. As a baseline, we consider Mistral-7B-Instruct-v0.2 used with the Langchain CSV agent, since it represents a typical open-source LLM with limited function-calling capabilities. Also, there is a significant gap between Claude 3.5 Sonnet and DeepSeek-R1-Distill-Qwen-1.5B while using the Llangraph CSV agent, which shows us that Claude's stronger reasoning abilities and performance make it far more effective for this task compared to DeepSeek.

While analyzing the models' responses, some interesting patterns emerged. For example, when asked about days of the week in one dataset, Claude returned months instead. Meanwhile, DeepSeek-R1-Distill-Qwen-1.5B struggled with naming variables correctly, frequently making coding mistakes.

V. Conclusion

This paper investigates how state-of-the-art AI frameworks and large language models (LLMs) handle structured tabular data from CSV files. The research demonstrates how advanced reasoning capabilities in LLMs work together with structured reasoning agents to enhance question-answering performance in real-world applications.

The evaluation assesses three large language models namely Claude 3.5 Sonnet, Mistral-7B-Instruct-v0.2, and DeepSeek-R1-Distill-Qwen-1.5B for their ability to reason and execute functions. The combination of Claude 3.5 Sonnet with the LangChain CSV agent produces the highest accuracy results when answering questions on tabular data. While open-weight models Mistral and DeepSeek demonstrate restricted capabilities in function calling and structured data interpretation, resulting in suboptimal performance.

References

[1] Kevin Bartley, "Big data statistics: How much data is there in the world?" 2024. [Online]. Available: https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/

[2] Aleksoft Editior Team, "Structured vs unstructured data explained with examples," 2024. [Online]. Available: https://www.altexsoft.com/blog/structured-unstructured-data//

[3] B. J. Juan Sequeda, Dean Allemang, "Increasing accuracy of llm-powered question answering on sql databases: Knowledge graphs to the rescue," 2024. [Online]. Available: http://sites.computer.org/debull/A24dec/p109.pdf

[4] A. Evangelatos, G. Filandrianos, and Maria, "Ails-ntua at semeval-2025 task 8: Language-to-code prompting and error fixing for tabular question answering," 2025. [Online]. Available: https://arxiv.org/abs/2503.00435

[5] J. Lovon, M. Mouysset, J. Oleiwan, J. G. Moreno, C. Damase-Michel, and L. Tamine, "Evaluating llm abilities to understand tabular electronic health records: A comprehensive study of patient data extraction and retrieval," 2025. [Online]. Available: https://arxiv.org/abs/2501.09384

[6] N. Chinaksorn and D. Wanvarie, "Llm-rag for financial question answering: A case study from set50," in 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2025.

[7] J.-P. Zhu, P. Cai, K. Xu, L. Li, Y. Sun, S. Zhou, H. Su, L. Tang, and Q. Liu, "Autotqa: Towards autonomous tabular question answering through multi-agent large language models," 2024. [Online]. Available: https://doi.org/10.14778/3685800.3685816

[8] Cubode Team, "How to use ai agents to analyze and process csv data: A comprehensive guide," 2024. [Online]. Available: https://medium.com/@cubode/comprehensive-guide-using-ai-agents-to-analyze-and-process-csv-data-a0259e2af761

[9] Anukriti Ranjan, "Langchain csv agent — a chain of function calls," 2023. [Online]. Available: https://anukriti-ranjan.medium.com/langchain-csv-agent-a-chain-of-function-calls-part-1-f0290e6c4bf3

[10] "Introduction to langchain," 2025. [Online]. Available: https://www.geeksforgeeks.org/introduction-to-langchain/

[11] "Agents," 2025. [Online]. Available: https://python.langchain.com/v0.1/docs/modules/agents/

[12] "Agents in artificial intelligence," 2025. [Online]. Available: https://www.geeksforgeeks.org/agents-artificial-intelligence/

[13] "Talk to your data using langchain csv agents and amazon bedrock," 2024. [Online]. Available: https://medium.com/@thallyscostalat/talk-to-your-data-using-langchain-csv-agents-and-amazon-bedrock-07ee3d35e9f7

[14] "Langgraph," 2025. [Online]. Available: https://langchain-ai.github.io/langgraph/#langgraphs-ecosystem

[15] H. Hoffman, "Langgraph: Build stateful ai agents in python," 2025. [Online]. Available: https://realpython.com/langgraph-python/

[16] "Claude 3.5 sonnet," 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet

[17] "Mistral-7b-instruct-v0.2," 2024. [Online]. Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[18] "Deepseek-r1-distill-qwen-1.5b," 2025. [Online]. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

[19] "Semeval 2025 - task 8: Databench, question-answering over tabular data," 2025. [Online]. Available: https://www.codabench.org/competitions/3360/

[20] "Evaluation method," 2025. [Online]. Available: https://github.com/jorses/databench_eval/blob/main/src/databench_eval/eval.py

Web Tracking Technologies - current research analysis

Resul Bedii Gümüş
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
Email: rbediigumus@gmail.com

Boban Joksimoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
Email:boban.joksimoski@finki.ukim.mk

***Abstract*—The rise of personalized online advertising has led to the proliferation of web tracking technologies that monitor users' online behavior across websites. While enabling targeted content delivery, these mechanisms – ranging from cookies to fingerprinting – raise serious privacy concerns due to their often opaque and persistent data collection practices. This paper presents a literature review of web tracking technologies, categorizing both traditional and emerging methods.**

***Index Terms*—Web tracking, browser fingerprinting, cookies, online privacy, user profiling, data protection, behavioral tracking.**

I. Introduction

The online advertising industry has grown exponentially, driven by the ability to deliver personalized content to users based on the vast amounts of data collected from their browsing behavior. This data collection is often facilitated through cookies and other tracking technologies, enabling advertisers and data brokers to profile individuals and serve targeted advertisements.

However, this practice raises significant privacy concerns. The data collected may persist throughout websites and be aggregated by third-party trackers or collected in a pool of data that may lead to personally identifiable information being revealed without users' consent. Moreover, recent tracking methodologies like browser and canvas fingerprints or tracking user behavior to construct a profile are gaining prominence over third-party cookies, which are harder to detect and opt out of.

In addition to laws and policies such as the General Data Protection Regulation (GDPR)[13] and private institution solution proposals such as Privacy Sandbox from Google[6], web users seek to control the data collected about them by utilizing privacy-enhancing technologies (PETs).

This study aims to introduce web tracking technologies by reviewing existing literature.

II. Web Tracking Technologies

A. Literature Review and Categorization

Understanding web tracking requires examining how key studies describe and analyze the technologies involved. Existing literature highlights the shift from basic mechanisms like cookies to more advanced and less detectable methods, such as fingerprinting. This section reviews foundational research to provide context on how tracking has evolved and its implications for user privacy.

The article *The Web Is Watching You: A Comprehensive Review of Web-Tracking Techniques and Countermeasures* [11] systematically categorizes and defines web tracking technologies based on their behavioral and technical characteristics. It adapts Roesner et al.'s taxonomy to classify tracking behavior into five categories: Analytics (Behavior A), Vanilla (Behavior B), Forced (Behavior C), Referred (Behavior D), and Personal (Behavior E). Each of these behaviors reflects different interactions between users, websites, and third-party trackers, ranging from basic analytics and cross-domain tracking using third-party cookies to more invasive methods such as forced redirections and direct personal interactions with tracker domains. Moreover, the article divides the tracking technologies into two main technical groups: stateful and stateless tracking. Stateful tracking methods store tracking information directly on users' devices and include commonly known techniques such as HTTP cookies, third-party cookies, Flash cookies (Local Shared Objects), HTML5 storage, and the JavaScript `window.name` property. Conversely, stateless tracking, also known as fingerprinting, uniquely identifies users by exploiting specific browser or system attributes without the need to store local data. Techniques under this category include JavaScript-based fingerprinting, font enumeration, canvas fingerprinting, visited-link style tracking, and JavaScript engine benchmarks. This categorization highlights significant differences in how tracking technologies operate and their implications for user privacy, particularly emphasizing that stateless tracking methods pose a more significant challenge to existing privacy protection measures due to their reliance on inherent device characteristics.

Building upon this foundational framework, the article *Web Tracking – A Literature Review on the State of Research* [4] employs a similar dual-layered approach to categorize web tracking technologies. Like the former, it adopts the Roesner taxonomy to describe tracking behaviors (A–E) while also presenting a detailed overview of tracking mechanisms. It expands on the landscape by introducing additional technologies such as Flash cookies (LSOs), cookie syncing, supercookies, and zombie cookies—each of which leverages diverse storage

mechanisms to ensure resilience against deletion and detection. Furthermore, it offers an enriched view of stateless tracking through advanced fingerprinting techniques like behavioral biometrics, which capture unique user traits such as typing rhythm or mouse movements.

In *Combating Web Tracking: Analyzing Web Tracking Technologies for User Privacy* [12], the authors analyze both the taxonomy of tracking technologies and their risk landscape. While the paper maintains the broad stateful versus stateless dichotomy, it introduces an extensive range of web tracking methods, including not only cookies, web beacons, and fingerprinting but also lesser-discussed mechanisms such as referrer headers, IP address-based tracking, and third-party tracking scripts. In addition to standard JavaScript-based techniques, the article details how scripts from analytics tools (e.g., Google Analytics, Hotjar) and third-party libraries silently collect data such as page visits, cursor movement, and time on page. It also examines how browser extensions and legacy browser plugins can function as potent tracking vectors, especially due to their privileged access to browser APIs and OS-level interactions. The article distinguishes these technologies not only by their storage behavior but also by their visibility to users, persistence, and potential for exploitation. Crucially, it dedicates significant attention to the security risks posed by tracking technologies: session hijacking through cookie theft, Cross-site request forgery (CSRF) via auto-included cookies, cross-site scripting (XSS) enabling unauthorized access to session identifiers, and the exploitation of referrer headers to leak private navigation data. It also flags the ability of malicious browser extensions and third-party scripts to collect and exfiltrate sensitive information, sometimes even bypassing sandboxing restrictions.

The article *Online Advertising: Analysis of Privacy Threats and Protection Approaches* [5] contextualizes tracking within the digital advertising ecosystem, offering a nuanced perspective centered around advertising infrastructure and its data collection practices. While the article does not adopt a formal behavioral taxonomy, it provides a functional categorization of tracking technologies used in online advertising, focusing on mechanisms that enable persistent, large-scale user profiling. These include cookie-based tracking and more covert methods such as fingerprinting, Flash cookies, canvas fingerprinting, and cookie matching. In particular, cookie matching is described as a sophisticated and privacy-invasive method that allows disparate advertising entities to synchronize identifiers and build cross-domain behavioral profiles.

The analysis extends into stateless techniques like browser and canvas fingerprinting, which uniquely identify users based on system-level attributes and rendering behavior and do not require any local storage, making them exceptionally difficult to detect or block. Flash cookies are noted for their resilience, as they can respawn deleted HTTP cookies and function across browsers. Beyond technical mechanisms, the article places a strong emphasis on the privacy threats posed by these technologies. These include the opaque aggregation of personal data by ad networks, persistent tracking that circumvents user deletion attempts, and the cross-site mapping of user behavior without meaningful consent. Importantly, the article highlights that such tracking practices often violate data protection regulations like the GDPR, particularly in contexts where personal identifiers or user activities are shared among ad brokers without transparency. This advertising-centric framing reinforces previously established tracking typologies and underscores the commercial incentives driving the development and deployment of increasingly evasive and persistent tracking technologies.

While much of the literature on web tracking focuses on conventional mechanisms such as cookies and fingerprinting, recent studies have also begun exploring more sophisticated and evasive tracking techniques. Among these, Real-Time Bidding (RTB) receives significant attention due to its central role in personalized advertising ecosystems. The RTB process, which involves auctioning ad space in real-time based on user attributes, is described in depth in both *Annoyed Users: Ads and AdBlock Usage in the Wild* [10] and *Online Advertising: Analysis of Privacy Threats and Protection Approaches* [5]. These studies highlight how advertisers receive detailed user data, such as IP addresses, device information, and behavioral history, before bidding on ad impressions, raising substantial privacy concerns.

In addition, WebRTC-based tracking is emerging as a potent stateless method for leaking IP addresses via Session Traversal Utilities for NAT (STUN) requests, even when users are behind VPNs or firewalls. This technique is flagged by *Evolution of Web Tracking Protection in Chrome* [9], which notes that fingerprinting detection tools are now being updated to catch WebRTC-based leaks, and by *Combating Web Tracking* [12], which lists WebRTC among the properties probed during fingerprinting audits.

The use of First-Party Cookie Tracking by Third-Parties is analyzed in *Cookie Swap Party: Abusing First Party Cookies for WebTracking* [2], which reveals how third-party scripts can manipulate and exfiltrate first-party cookies. This practice circumvents traditional third-party cookie blockers and allows for persistent cross-site user tracking, undermining many browser-level privacy defenses.

Although less frequently discussed, Content Delivery Network (CDN) Tracking is also highlighted as a growing concern. CDNs, while traditionally used to optimize content delivery, are increasingly being leveraged—intentionally or otherwise—for tracking purposes. This is especially evident in RTB pipelines, where ad creatives and tracking scripts are served from CDN domains, as observed in *Annoyed Users* [10] and *Effective Techniques for Protecting the Privacy of Web Users* [1].

B. Technologies

This section provides an overview of individual web tracking technologies, both traditional and emerging. Each subsection briefly introduces a specific technique, explaining how it works and why it raises privacy concerns. The goal is to present an understanding of how these mechanisms operate,

how they are used in practice (often in combination), and what makes them particularly difficult to detect or prevent.

1) HTTP Cookies – Session, Persistent, Third-Party: HTTP cookies are small data files that websites store on a user's browser to manage session state and remember user preferences. They come in three types: session cookies, which are temporary and deleted when the browser closes; persistent cookies, which remain on the device for a set time and enable long-term tracking; and third-party cookies, which are set by domains other than the one the user is visiting and are commonly used for cross-site tracking in advertising and analytics. Third-party cookies can track users over many websites, building detailed behavioral profiles [11, 3].

2) Flash Cookies (LSOs): Flash cookies, also known as Local Shared Objects (LSOs), are stored using Adobe Flash rather than the browser's cookie system. They can store significantly more data (up to 100KB), do not expire by default, and are not easily managed through traditional browser privacy settings. These cookies are particularly invasive because they can be used to regenerate previously deleted HTTP cookies, a practice known as zombie cookie behavior. Their persistence across sessions and even browsers makes them especially difficult to remove [11, 5]. Thankfully, with the end-of-life of Flash Player in 2020, the exploitation of this attack vector is not as beneficial as it used to be.

3) HTML5 Web Storage: HTML5 introduced Web Storage APIs—`localStorage` and `sessionStorage`—as a more flexible and high-capacity alternative to cookies. These APIs allow websites to store key-value data directly in the browser, supporting larger volumes of information and better performance. Because the data stored is not tied to HTTP requests, it bypasses traditional cookie management and blocking mechanisms. This makes Web Storage attractive for persistent tracking and is sometimes exploited to recreate deleted cookies [9, 5].

4) First-Party Cookie Tracking by Third-Parties: This tracking method involves third-party scripts operating within a first-party context using the `document.cookie` API to set cookies. Although technically classified as first-party cookies, these can be read and exfiltrated by the third-party script, effectively bypassing browser policies that restrict third-party cookies. This enables hidden cross-site tracking and identity linkage across domains, even when third-party cookies are blocked. It is commonly used in analytics services like Google Analytics [11, 12].

5) Cookie Syncing / Cookie Matching: Cookie syncing (or cookie matching) enables multiple advertising entities to align their identifiers for a user. Since browsers enforce the Same-Origin Policy, direct sharing of cookies between third-party domains is restricted. To bypass this, trackers initiate redirects or embed shared identifiers in URLs to synchronize user IDs. This process allows advertisers to build unified cross-site profiles and powers data exchange in real-time bidding (RTB) environments. The practice raises significant privacy concerns as it expands user tracking without explicit consent [11, 5].

6) Zombie Cookies: Zombie cookies are tracking identifiers that are automatically regenerated after deletion. They achieve this by using multiple redundant storage locations such as Flash cookies, HTML5 storage, and ETag headers. These mechanisms allow the original identifier to be restored even after the user attempts to clear it. This kind of tracking subverts user expectations and privacy settings, making it one of the most persistent and controversial methods of surveillance [11].

7) Supercookies: Supercookies are tracking techniques that do not rely on traditional client-side storage like HTTP cookies. They include mechanisms such as ISP-injected identifiers, ETag-based tracking, or advanced fingerprinting techniques that uniquely identify a user without storing any data on the user's device. These techniques are challenging to detect and delete and often survive browser resets. As such, supercookies are considered highly invasive and difficult to block using conventional privacy tools [5].

8) JavaScript-Based Fingerprinting: JavaScript-based fingerprinting involves collecting a combination of browser-specific attributes using client-side JavaScript. This includes properties such as browser version, installed plugins, MIME types, screen resolution, color depth, and system fonts. These attributes, when combined, create a highly unique identifier for each device. Early studies, such as Panopticlick, demonstrated the effectiveness of this method in distinguishing users without storing any state on the client side. Trackers exploit JavaScript APIs like `navigator`, `screen`, and `window` to build these fingerprints. The uniqueness of this information, especially when paired with details like time zones or HTTP headers, enables persistent user identification even across private browsing modes or after cookie deletion [11, 4, 12].

9) Canvas Fingerprinting: Canvas fingerprinting utilizes the HTML5 `canvas` element to generate a unique device identifier. This is done by rendering text or WebGL graphics on the canvas and then hashing the resulting pixel values. The technique leverages subtle differences in how operating systems and hardware render text and images—differences arising from factors like font rendering, anti-aliasing, and subpixel smoothing. Since this method doesn't rely on local storage, it's difficult for traditional anti-tracking tools to detect or block. Several studies found that canvas fingerprinting is widespread among advertising and analytics domains, where it operates silently without user interaction [11, 5, 12].

10) Font Enumeration: Font enumeration is a fingerprinting method that determines which fonts are installed on a user's system. This is often achieved using JavaScript to measure the rendering differences of text displayed in different fonts. By iterating through a list of commonly available fonts and measuring the rendered dimensions or visual differences, trackers can compile a list of active fonts on a machine. Since font libraries differ across devices and regions, this list becomes a strong entropy signal in building a unique fingerprint. This method is especially effective when combined with other fingerprinting vectors like screen resolution and browser version [11, 12].

11) Behavioral Biometrics: Behavioral biometrics refers to tracking techniques that analyze users' behavioral patterns, such as typing speed, mouse movement, scroll dynamics, and click behavior, to create persistent user profiles. These techniques are considered stateless because they don't depend on cookies or local storage; rather, they extract behavioral features during user interactions. As machine learning models evolve, they can identify users with increasing accuracy based on these subtle cues. Behavioral fingerprinting has also been proposed as a means to differentiate bots from humans in fraud detection and bot prevention scenarios. Though less commonly deployed than traditional fingerprinting, it poses significant privacy challenges due to its covert nature and profiling capabilities [12].

12) WebRTC-Based Tracking: WebRTC-based tracking exploits the peer-to-peer communication capabilities of the WebRTC API. Specifically, by initiating a STUN (Session Traversal Utilities for NAT) request, websites can retrieve both the public and local IP addresses of a device. Even behind a VPN, the local IP address may still be exposed, thus undermining anonymization efforts. Unlike other fingerprinting techniques, WebRTC tracking offers geographic and network-level identifiers, making it highly attractive for surveillance and profiling. The subtlety of this tracking method—often executed silently in the background—makes it hard to detect without specialized privacy tools or browser-level mitigations [12, 9].

13) IP Address-Based Tracking: IP-based tracking relies on using a user's unique IP address, which is assigned each time they connect to the internet. This identifier allows websites and trackers to infer a user's approximate geographic location, Internet Service Provider (ISP), and network configuration. It is often employed to distinguish between users or devices, especially in server logs or for delivering region-specific content. While useful, the accuracy is limited due to shared or dynamic IPs, commonly assigned by ISPs or used in shared networks like corporate environments or public Wi-Fi [12].

14) Referrer Header Tracking: This technique uses the HTTP `Referer` header to track how users navigate from one webpage to another. When a user clicks a link, the browser automatically sends the URL of the originating page to the destination server. This enables web operators to monitor user flow, determine entry points, and optimize marketing strategies. However, referrer headers may inadvertently leak sensitive information like search terms or private URLs. Modern browsers attempt to mitigate these risks through features like the `Referrer-Policy` header, with Chrome adopting stricter defaults in recent versions [12, 9].

15) Web Beacons / Pixel Tags: Web beacons—also known as tracking pixels—are tiny, transparent $1x1$ pixel images embedded in webpages or emails. When loaded, they make a request to the server hosting the image, transmitting data such as the user's IP address, browser type, and timestamp. In emails, they signal when a message has been opened and how many times. Because beacons are visually invisible and often bundled in third-party scripts or content, they are a covert yet powerful tool for behavior tracking across websites and platforms [12].

16) Real-Time Bidding (RTB) Tracking: RTB is a central mechanism in programmatic advertising, where ad space is auctioned off in real-time as a user loads a webpage. Ad exchanges collect detailed information about the user, such as device details, browsing history, IP address, and location, and then send this data to multiple advertisers, who place bids to serve an ad. This process typically completes within milliseconds. RTB allows highly granular user targeting and personalization but also results in the widespread dissemination of personal data to multiple, often opaque, parties [5, 10].

17) Content Delivery Network (CDN) Tracking: CDNs are infrastructure providers that cache and deliver content from geographically distributed servers to improve load times. However, they can also inadvertently or deliberately be involved in user tracking. When ad content or scripts are served from a CDN domain, repeated access patterns can be monitored. Some trackers co-locate with CDNs or utilize CDN-provided analytics features. Studies show CDNs like Akamai and SoftLayer frequently appear in ad delivery chains, serving both functional content and trackers, thus blurring the line between performance optimization and surveillance [10, 5].

III. Future Research

This article has reviewed web tracking technologies, highlighting their technical mechanisms and privacy implications. While privacy-enhancing technologies (PETs) play a critical role in countering these techniques, their evaluation falls outside the scope of this paper and will be addressed in future work. Understanding tracking systems in detail will be a necessary first step toward assessing the effectiveness of PETs and informing the development of more robust privacy solutions.

Building on prior studies that either use large-scale automated testing [1, 7, 8] or analyze user behavior [10], the next stage of this research will systematically test PETs through automated browsing experiments. Platforms like OpenWPM and FP-Crawler offer reliable means to measure how these tools affect tracking activity [12].

The study will classify PETs by their defenses against specific tracking technologies and evaluate their impact through controlled experiments. Additionally, it will incorporate location-based testing to examine how PET effectiveness varies. By simulating browsing from different geographic regions, this work will provide new insights into the influence of location on web tracking and privacy protection.

Acknowledgments

This work was supported partially by the European Union in the framework of ERASMUS MUNDUS, Project CyberMACS (Project #101082683) (https://cybermacs.eu).

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje.

References

[1] Maryam Bubukayr and Mounir Frikha. "Effective Techniques for Protecting the Privacy of Web Users". In: *Applied Sciences* 13.5 (2023), p. 3191. DOI: 10.3390/app13053191.

[2] Quan Chen et al. "Cookie swap party: Abusing first-party cookies for web tracking". In: *Proceedings of the Web Conference 2021*. 2021, pp. 2117–2129.

[3] Nadr El Hana, Maria Mercanti-Guérin, and Ouidade Sabri. "Cookiepocalypse: What Are the Most Effective Strategies for Advertisers to Reshape the Future of Display Advertising?" In: *Technological Forecasting and Social Change* 188 (2023), p. 122297. DOI: 10.1016/j.techfore.2022.122297.

[4] Tatiana Ermakova et al. "Web Tracking – A Literature Review on the State of Research". In: (2018).

[5] José Estrada-Jiménez et al. "Online Advertising: Analysis of Privacy Threats and Protection Approaches". In: *Computer Communications* 100 (2017), pp. 32–51. DOI: 10.1016/j.comcom.2016.12.016.

[6] Minjun Long and David Evans. "Evaluating Google's Protected Audience Protocol". In: *Proceedings on Privacy Enhancing Technologies* 2024.4 (2024), pp. 892–906. DOI: 10.56553/popets-2024-0147.

[7] Siddharth M. Madikeri and Vijay K. Madisetti. "Ad Blockers & Online Privacy: A Comparative Analysis of Privacy Enhancing Technologies (PET)". In: *Journal of Software Engineering and Applications* 17.5 (2024), pp. 378–395. DOI: 10.4236/jsea.2024.175021.

[8] Johan Mazel, Richard Garnier, and Kensuke Fukuda. "A Comparison of Web Privacy Protection Techniques". In: *Computer Communications* 144 (2019), pp. 162–174. DOI: 10.1016/j.comcom.2019.04.005.

[9] Ronghao Pan and Antonio Ruiz-Martínez. "Evolution of Web Tracking Protection in Chrome". In: *Journal of Information Security and Applications* 79 (2023), p. 103643. DOI: 10.1016/j.jisa.2023.103643.

[10] Enric Pujol, Oliver Hohlfeld, and Anja Feldmann. "Annoyed Users: Ads and Ad-Block Usage in the Wild". In: *Proceedings of the 2015 Internet Measurement Conference*. 2015, pp. 93–106. DOI: 10.1145/2815675.2815705.

[11] Iskander Sanchez-Rola et al. "The Web Is Watching You: A Comprehensive Review of Web-Tracking Techniques and Countermeasures". In: *Logic Journal of IGPL* 25.1 (2017), pp. 18–29. DOI: 10.1093/jigpal/jzw041.

[12] Kyungmin Sim, Honyeong Heo, and Haehyun Cho. "Combating Web Tracking: Analyzing Web Tracking Technologies for User Privacy". In: *Future Internet* 16.10 (2024), p. 363. DOI: 10.3390/fi16100363.

[13] Peter Snyder and Chris Kanich. "Measuring Cookies and Web Privacy in a Post-GDPR World". In: *Proceedings of the Internet Measurement Conference* (2015).

Python for Mass Spectrometry: A Comprehensive Overview of Libraries for Data Processing and Analysis

1st Elena Atanasoska
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
elena.atanasoska@finki.ukim.mk

2th Riste Stojanov
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
riste.stojanov@finki.ukim.mk

***Abstract*—Mass spectrometry is a powerful analytical technique that is used across various fields, including biology, chemistry, and medicine. The increasing complexity of mass spectrometry data requires efficient computational tools for processing and analysis. Python, with its extensive libraries and community support, has become as a leading platform mass spectrometry data handling. This overview provides a comprehensive survey of Python libraries specifically designed for mass spectrometry data processing and analysis, such as pyOpenMS, Pyteomics and matchms. These libraries facilitate tasks ranging from data import and cleaning to peak detection and quantification, enabling researchers to extract meaningful insights from mass spectrometry datasets efficiently. Using these tools, scientists can streamline workflows, improve data quality, and accelerate discovery in fields such as proteomics and metabolomics. This review aims to guide researchers in selecting the most appropriate Python libraries for their mass spectrometry data analysis needs, highlighting best practices and future directions in the integration of Python with mass spectrometry research.**

***Index Terms*—Mass Spectrometry, Python Libraries, Data Processing, Data Analysis, Mass Spectrometry Imaging, Open-Source Software, Computational Tools, Proteomics, Metabolomics**

I. Introduction

Mass spectrometry (MS) is an *analytical technique* used to identify chemical substances by sorting gaseous ions based on the ratio of their **mass to charge (m/z)** [1, 2, 3]. This method allows for the determination of the **molecular weight** of particles and provides both *qualitative* information about the structure of a molecule and *quantitative* data on its concentration within a sample [4]. The fundamental principle of MS involves *ionizing* the molecules present in a sample and subsequently separating the resulting ions according to their **mass-to-charge ratio**. The outcome of this process is typically visualized as a *mass spectrum*, which is a graphical representation plotting the **relative abundance** of the detected ions against their corresponding **m/z values** [5, 6].

This process typically involves several steps, including *ionization*, *acceleration*, *deflection*, *detection*, and *data processing* [7]. Different ionization techniques, such as **Electron Ionization (EI)** and **Chemical Ionization (CI)**, are used depending on the nature of the analyte [8]. Mass analyzers, like *time-of-flight (TOF)* and *quadrupole mass analyzers*, separate ions based on their mass [9, 10]. The resulting mass spectra provide valuable information about the molecules present in the sample.

The analysis of complex data generated by mass spectrometers often requires *sophisticated computational tools* [11]. **Python** has become increasingly popular for scientific data analysis due to its *simplicity*, *versatility*, and the availability of numerous powerful libraries. Its intuitive syntax and readability make it easier to learn and use compared to some other programming languages. Furthermore, Python's *open-source* nature and large, active community ensure extensive support, documentation, and a continuously growing collection of libraries for various tasks, including *data manipulation*, *statistical analysis*, *machine learning*, and *visualization* [12, 13].

II. Python for Mass Spectrometry Data Processing and Analysis

A. Why Python?

Python offers several strong advantages for mass spectrometry data processing and analysis. Its **simplicity** and **readability** make it accessible to researchers with varying levels of programming experience, reducing the time spent on learning the language and allowing for a greater focus on *data analysis* [14]. Python's extensive collection of **free** and **well-supported** libraries is a significant benefit. Libraries such as *NumPy* and *Pandas* provide efficient data structures and functions for numerical computations and data manipulation, which are essential for handling large mass spectrometry datasets. *Matplotlib* and *Seaborn* offer powerful tools for creating informative visualizations of mass spectra and analytical results.

Python's **versatility** allows it to be used for a wide range of tasks, from developing web applications and APIs to implementing complex scientific algorithms [12]. Its **cross-platform compatibility** ensures that code can be run on different operating systems, helping collaboration and deployment. Moreover, Python's strong support for *machine learning* through libraries

like *Scikit-learn* makes it ideal for advanced data interpretation and pattern recognition in mass spectrometry data [15].

B. Commonly Used Python Libraries

Several Python libraries are commonly employed for the initial stages of mass spectrometry data processing. These include:

- **Pandas**: Provides efficient *data structures*, such as *DataFrames*, which are essential for organizing and analyzing **tabular data**, including mass spectrometry results [16].
- **NumPy**: A core library for **numerical computing**, offering support for *multi-dimensional arrays* and advanced mathematical functions, which are crucial for processing **intensity values** and *m/z ratios* [17].
- **Matplotlib**: A widely used library for generating **static, interactive, and animated visualizations**, allowing researchers to create plots of *mass spectra*, *chromatograms*, and other relevant data [18].
- **Seaborn**: Built on top of Matplotlib, it provides a **higher-level interface** for creating *statistically-oriented* and *visually appealing* plots, improving the interpretability of mass spectrometry data [18].

C. Advanced Python Libraries

For more specialized tasks in mass spectrometry data analysis, several dedicated Python libraries are available. These include:

- **pyOpenMS**: An *open-source* library offering a comprehensive suite of functionalities for computational **mass spectrometry**, particularly in **proteomics** and **metabolomics**. It includes tools for **file handling** (e.g., mzXML, mzML), **signal processing** (smoothing, filtering, peak picking), **identification analysis** (peptide search, PTM analysis), and **quantitative analysis** (label-free, SILAC) [19].
- **Pyteomics**: A lightweight and modular library designed for handling **proteomics data**, providing tools for *polypeptide property calculation*, **proteomics file format access**, and *peptide sequence manipulation*. It supports various file formats and offers functionalities tailored for **common proteomics analysis tasks** [20].
- **matchms**: A specialized library focused on **processing and comparing** tandem mass spectrometry (**MS/MS**) data. It enables *importing, cleaning*, and **computing spectral similarities** using various similarity measures. Additionally, it supports multiple *spectral data formats*, enhancing its utility in spectral analysis workflows [21].

III. pyOpenMS: A Versatile Toolkit for Mass Spectrometry Data Analysis

pyOpenMS is an *open-source* Python library that serves as an interface to the **OpenMS** C++ library, a well-established framework for **computational mass spectrometry** [22]. This Python binding provides researchers with seamless access to a **comprehensive set of algorithms** specifically designed for **proteomics** and **metabolomics** data analysis.

The library is **cross-platform**, supporting *Windows, Linux*, and *macOS*, making it a versatile tool for a wide range of users [23]. By leveraging its **C++ foundation**, **pyOpenMS** ensures efficient execution of computationally intensive tasks while maintaining the flexibility of Python. The Python interface further facilitates **rapid prototyping** and enables seamless integration with other **scientific Python libraries** [19].

A. Functionalities Relevant to Mass Spectrometry

pyOpenMS encompasses a wide range of functionalities essential for mass spectrometry data analysis. These can be broadly categorized into file handling, signal processing, peak picking, identification analysis, and quantitative analysis:

- **File Handling** - The **pyOpenMS** library provides efficient file handling capabilities, supporting a wide range of mass spectrometry formats, including *mzXML*, *mzML*, *TraML*, *mzTab*, *FASTA*, and *PepXML*. It also manages quantitative analysis formats such as *featureXML* and *consensusXML*, alongside standards like *mzML* and *mzIdentML*, developed by the Proteomics Standards Initiative (PSI) [24, 25]. Furthermore, it supports the emerging *mzQuantML* standard.
  Data is processed in **NumPy** arrays, allowing seamless integration with scientific Python tools such as **Matplotlib** and **SciPy**. With support for over **30 file formats**, including *PepXML* and *ProtXML*, **pyOpenMS** is highly versatile. For handling large *mzML* files, the library provides multiple optimization techniques, including *indexing*, *streaming*, and *caching*, ensuring adaptability to different workflows and analytical needs [19].
- **Signal Processing** - **pyOpenMS** includes a suite of signal processing tools designed to enhance mass spectrometry data quality. These include *data smoothing* to reduce noise, *filters* to improve signal clarity, and *de-isotoping* to resolve isotopic envelopes. The library also offers *retention time correction* algorithms to address chromatographic variations and methods for *chromatogram extraction* to analyze elution profiles.
  For high-resolution data, **pyOpenMS** provides **Time-of-Flight (TOF) calibration** for precise mass measurements. These signal processing capabilities are crucial for preparing raw data, refining spectral accuracy, and improving downstream analyses, such as *peak picking* and *compound identification* [19].
- **Peak Picking** - Peak picking is an essential step in mass spectrometry, involving the identification and extraction of relevant spectral signals. **pyOpenMS** offers multiple algorithms tailored for different data types:
  - **PeakPickerHiRes** – Optimized for high-resolution data, this algorithm converts *profile-mode spectra* into a condensed *centroid mode*, where each peak is represented by a single *mass-to-charge (m/z)* and intensity value.

- **FeatureFindingMetabo** – Used in *metabolomics*, this method assembles mass spectrometry signals into distinct features representing individual metabolites.
- **FeatureFinderCentroided** and **FeatureFinderSuperHirn** – Designed for *liquid chromatography-mass spectrometry (LC-MS/MS)*, these algorithms detect *two-dimensional features* by considering both *m/z* and *retention time*.

The library also includes additional feature detection approaches suited for **proteomics**, **untargeted metabolomics**, and **targeted metabolomics** experiments. These diverse peak-picking algorithms ensure that **pyOpenMS** can effectively handle various types of mass spectrometry data and experimental workflows [19].

IV. Pyteomics: A Versatile Library for Proteomics Data Handling

Pyteomics is a cross-platform, open-source Python library designed to provide a comprehensive suite of tools for handling proteomics data. With its **lightweight** and **modular** architecture, Pyteomics enables efficient execution of key tasks, including *polypeptide property calculations*, *proteomics data format parsing*, and *peptide sequence manipulation*.

A key strength of Pyteomics is its **modularity**, allowing researchers to utilize only the necessary components, thereby optimizing computational efficiency and seamless integration into custom workflows. The library supports widely used proteomics data formats, ensuring compatibility with existing tools and pipelines.

Pyteomics aims to provide a **versatile, reliable, and well-documented** resource for the proteomics community. By leveraging Python's growing role in scientific programming, it facilitates **reproducible statistical data analysis** and **rapid software prototyping**, making it an essential tool for modern proteomics research [20].

A. Key Features and Tools

Pyteomics offers a range of modules that cater to different aspects of proteomics data handling. These include tools for polypeptide property calculation, accessing various proteomics data formats, and manipulating peptide sequences:

- **Polypeptide Property Calculation**
  Pyteomics provides several modules for computing key physicochemical properties of polypeptides. The `pyteomics.mass` module enables the calculation of monoisotopic and average masses for peptides, proteins, modified peptides, and peptide ions. It also estimates isotopic distribution abundances and determines the most probable isotopic state of a molecule.
  The `calculate_mass()` function supports polypeptide sequences, chemical formulas, and elemental compositions while computing mass-to-charge (m/z) ratios for specific ion types and charge states. For charge and isoelectric point (pI) prediction, the `pyteomics.electrochem` module provides functions like `charge()` and `pI()`, leveraging the Henderson-Hasselbalch equation [26].
  Additionally, Pyteomics includes tools for predicting chromatographic retention times. The `pyteomics.achrom` module implements additive models for calibrating retention time predictions based on experimental data. For a more detailed approach, the `pyteomics.biolccc` module incorporates the BioLCCC model, estimating the effect of experimental parameters on peptide separation [27].
  These comprehensive tools support various proteomics applications, from peptide identification to optimizing separation methods [28].
- **Proteomics Data Format Access**
  Pyteomics provides extensive support for proteomics data formats, ensuring seamless integration with experimental workflows and analysis pipelines.
  For raw mass spectrometry data, the library includes modules for mzML (`pyteomics.mzml`, `pyteomics.mzmlb`) and mzXML (`pyteomics.mzxml`), allowing users to iterate through spectra and extract spectrum-specific details. It also supports the Mascot Generic Format (MGF) through the `pyteomics.mgf` module, offering classes like `MGF` and `IndexedMGF` for structured data access.
  Pyteomics handles search engine outputs via modules for pepXML (`pyteomics.pepxml`), mzIdentML (`pyteomics.mzid`), protXML (`pyteomics.protxml`), and X! Tandem output (`pyteomics.tandem`). Additional modules support FASTA databases (`pyteomics.fasta`, `pyteomics.peff`), mzTab files (`pyteomics.mztab`), and formats like featureXML, trafoXML, idXML, and TraML.
  Notably, Pyteomics includes the `ppx` package, which enables programmatic access to ProteomeXchange repositories such as PRIDE and MassIVE [29]. The `read()` function serves as a unified entry point for reading multiple data formats, simplifying data acquisition and manipulation [20].
- **Peptide Sequence Manipulation**
  The `pyteomics.parser` module provides a suite of functions for handling peptide sequences, including those with modifications.
  Pyteomics adopts a custom IUPAC-derived peptide sequence notation called modX, which represents modified amino acid residues and terminal groups. The `parse()` function converts modX sequences into lists of amino acid residues for further analysis.
  For in silico digestion, the `cleave()` and `xcleave()` functions simulate enzymatic cleavage based on specified rules, such as trypsin specificity, while preserving original sequence indices. The `isoforms()` function generates all possible modified peptide sequences given a set of variable modifications.

Additional utilities include `length()` for computing peptide sequence length in modX format and `amino_acid_composition()` for determining amino acid composition.
These tools are crucial for proteomics workflows, aiding in theoretical spectrum generation, enzymatic digestion simulation, and post-translational modification analysis [20].

V. matchms: Streamlining Tandem Mass Spectrometry Data Processing and Comparison

matchms is an open-access Python package specifically developed for the processing, cleaning, and comparison of tandem mass spectrometry (MS/MS) data. Designed to facilitate the implementation of reproducible workflows, ***matchms*** enables researchers to transform raw MS/MS data from common file formats into pre- and post-processed spectral data and to perform large-scale spectral similarity comparisons.

The library supports a wide range of popular spectral data formats and offers an array of tools for metadata cleaning, peak filtering, and the application of various pairwise similarity measures. Its extensibility allows users to integrate custom similarity measures, and it includes efficient methods for handling large datasets. By focusing specifically on MS/MS data, ***matchms*** provides researchers with a dedicated toolkit for analyzing fragmentation spectra, which are crucial for compound identification and structural elucidation [21].

A. Functionalities for Processing and Comparing MS/MS Data

matchms provides a comprehensive suite of functionalities tailored to the unique characteristics of tandem mass spectrometry data. These include capabilities for importing data from various formats, cleaning and filtering spectra, and calculating spectral similarities using a variety of metrics.

- ***Importing MS/MS Data***
  matchms supports the import of MS/MS data from several common formats, including mzML, mzXML, msp, metabolomics-USI, MGF, and JSON files (such as those generated by GNPS). The library leverages **Pyteomics** for importing mzML, mzXML, and MGF files, indicating a collaborative approach within the Python mass spectrometry community.
  matchms also offers a `Pipeline` class, which simplifies the creation and management of complex workflows involving data import, processing, and score computations. This class allows users to define and execute a series of steps in a structured and reproducible manner, enhancing the accessibility of typical ***matchms*** workflows [30].
  The ability to import data from a diverse range of formats ensures that ***matchms*** can be readily integrated into various research settings and can handle data from different mass spectrometry instruments and software platforms [21]. Furthermore, its compatibility with widely used spectral repositories and open data standards facilitates data sharing and interoperability within the scientific community.
- ***Cleaning MS/MS Spectra***
  A crucial aspect of working with MS/MS data is ensuring its quality and consistency. ***matchms*** provides a variety of filters for metadata cleaning, validation, and basic peak filtering. These filters can harmonize metadata fields, derive missing information from other metadata fields, and repair or remove incorrect annotations.
  Many of the provided metadata cleaning filters are specifically designed to handle and improve GNPS-style MGF or JSON datasets, which are widely used in the metabolomics community. ***matchms*** also includes a dedicated library cleaning pipeline aimed at curating and correcting annotated mass spectral libraries, enhancing the quality of data used for library searching and machine learning model training.
  For peak filtering, ***matchms*** offers functions to select peaks based on m/z range or relative intensity, allowing users to remove noise and focus on the most significant signals. These comprehensive cleaning and filtering capabilities are essential for improving the accuracy and reliability of downstream spectral comparisons and analyses [30].
- ***Calculating Spectral Similarity***
  matchms is particularly strong in its ability to calculate spectral similarities between MS/MS spectra. The library offers a range of pairwise similarity measures, including the widely used `CosineGreedy` and `ModifiedCosine` scores, as well as other metrics like `NeutralLossesCosine` and `FingerprintSimilarity`. The `CosineGreedy` score, for example, calculates the cosine similarity between two spectra by finding the best possible matches between their peaks within a given tolerance.
  For molecules with known structures (e.g., represented by SMILES or InChI keys in the metadata), ***matchms*** can calculate similarity based on molecular fingerprints using the `rdkit` library. A key strength of ***matchms*** is its extensibility, allowing users to integrate custom similarity measures. Notable examples of spectrum similarity measures tailored for ***matchms*** include `Spec2Vec` and `MS2DeepScore`, which utilize machine learning approaches for spectral comparison.
  To enhance efficiency when comparing large numbers of spectra, ***matchms*** can utilize faster similarity measures for initial pre-selection and supports storing results in sparse data formats, reducing memory usage and computational time. These diverse spectral similarity measures and efficiency optimizations make ***matchms*** a powerful tool for tasks such as spectral library searching, compound annotation, and molecular networking [21].

VI. Conclusion

The landscape of mass spectrometry data analysis is increasingly shaped by powerful and versatile Python libraries.

pyOpenMS stands out as a comprehensive toolkit, offering a wide range of functionalities from raw data processing to sophisticated analytical methods, making it suitable for both proteomics and metabolomics research. Its ability to handle multiple file formats and perform complex analyses makes it a key tool in the field. **Pyteomics**, with its focus on proteomics, provides a modular and efficient set of tools for calculating polypeptide properties, accessing various data formats, and manipulating peptide sequences. Its lightweight design and specific modules are made for the needs of the proteomics community. **matchms**, dedicated to tandem mass spectrometry data, excels in importing, cleaning, and comparing MS/MS spectra. Its specialized tools for spectral similarity calculations and library cleaning address the unique challenges of analyzing fragmentation data.

In conclusion, these three Python libraries provide a strong set of tools for mass spectrometry data analysis. Their development reflects the growing importance of Python in scientific research and provides the community with open-source, flexible tools for tackling the complexities of modern mass spectrometry experiments. By understanding the unique capabilities of each library, researchers can build tailored workflows to address their specific research questions, ultimately advancing our understanding of biological systems through mass spectrometry.

References

[1] E. De Hoffmann and V. Stroobant, *Mass spectrometry: principles and applications*. John Wiley & Sons, 2007.

[2] A. El-Aneed, A. Cohen, and J. Banoub, "Mass spectrometry, review of the basics: electrospray, maldi, and commonly used mass analyzers," *Applied spectroscopy reviews*, vol. 44, no. 3, pp. 210–230, 2009.

[3] H. Awad, M. M. Khamis, and A. El-Aneed, "Mass spectrometry, review of the basics: ionization," *Applied spectroscopy reviews*, vol. 50, no. 2, pp. 158–175, 2015.

[4] S. Maher, F. P. Jjunju, and S. Taylor, "Colloquium: 100 years of mass spectrometry: Perspectives and future trends," *Reviews of Modern Physics*, vol. 87, no. 1, pp. 113–135, 2015.

[5] K. K. Murray, R. K. Boyd, M. N. Eberlin, G. J. Langley, L. Li, and Y. Naito, "Definitions of terms relating to mass spectrometry (iupac recommendations 2013)," *Pure and Applied Chemistry*, vol. 85, no. 7, 2013.

[6] J. Koomen, D. Hawke, and R. Kobayashi, "Developing an understanding of proteomics: an introduction to biological mass spectrometry," *Cancer investigation*, vol. 23, no. 1, pp. 47–59, 2005.

[7] C. S. Chu and C. B. Lebrilla, "Introduction to modern techniques in mass spectrometry," in *Biomedical Applications of Biophysics*, pp. 137–154, Springer, 2010.

[8] K. J. Margolin Eren, O. Elkabets, and A. Amirav, "A comparison of electron ionization mass spectra obtained at 70 ev, low electron energies, and with cold ei and their nist library identification probabilities," *Journal of Mass Spectrometry*, vol. 55, no. 12, p. e4646, 2020.

[9] N. Mirsaleh-Kohan, W. D. Robertson, and R. N. Compton, "Electron ionization time-of-flight mass spectrometry: Historical review and current applications," *Mass spectrometry reviews*, vol. 27, no. 3, 2008.

[10] M. S. Munson and F.-H. Field, "Chemical ionization mass spectrometry. i. general introduction," *Journal of the American Chemical society*, vol. 88, no. 12, 1966.

[11] C. Hoang, W. Uritboonthai, L. Hoang, E. M. Billings, A. Aisporna, F. A. Nia, R. J. Derks, J. R. Williamson, M. Giera, and G. Siuzdak, "Tandem mass spectrometry across platforms," *Analytical Chemistry*, vol. 96, no. 14, pp. 5478–5488, 2024.

[12] I. Stančin and A. Jović, "An overview and comparison of free python libraries for data mining and big data analysis," in *2019 42nd International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 977–982, IEEE, 2019.

[13] H. Fangohr, "A comparison of c, matlab, and python as teaching languages in engineering," in *Computational Science-ICCS 2004: 4th International Conference, Kraków, Poland, June 6-9, 2004, Proceedings, Part IV 4*, pp. 1210–1217, Springer, 2004.

[14] S. Bassi, "A primer on python for life science researchers," *PLoS computational biology*, vol. 3, no. 11, p. e199, 2007.

[15] Snowflake, "Why data scientists choose python for machine learning and artificial intelligence," March 25 2025. Accessed March 25, 2025. Available from: https://www.snowflake.com/trending/why-data-scientists-choose-python-ml-and-ai/.

[16] W. McKinney *et al.*, "pandas: a foundational python library for data analysis and statistics," *Python for high performance and scientific computing*, vol. 14, no. 9, pp. 1–9, 2011.

[17] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[18] E. Bisong and E. Bisong, "Matplotlib and seaborn," *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, pp. 151–165, 2019.

[19] H. L. Röst, U. Schmitt, R. Aebersold, and L. Malmström, "pyopenms: a python-based interface to the openms mass-spectrometry algorithm library," *Proteomics*, vol. 14, no. 1, pp. 74–77, 2014.

[20] L. I. Levitsky, J. A. Klein, M. V. Ivanov, and M. V. Gorshkov, "Pyteomics 4.0: five years of development of a python proteomics framework," *Journal of proteome research*, vol. 18, no. 2, pp. 709–714, 2018.

[21] F. Huber, S. Verhoeven, C. Meijer, H. Spreeuw, E. M. V. Castilla, C. Geng, J. J. van der Hooft, S. Rogers, A. Belloum, F. Diblen, *et al.*, "Matchms-processing and similarity evaluation of mass spectrometry data," *BioRxiv*, pp. 2020–08, 2020.

[22] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hus-

song, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, *et al.*, "Openms–an open-source software framework for mass spectrometry," *BMC bioinformatics*, vol. 9, pp. 1–11, 2008.

[23] pyOpenMS, "Summary — pyopenms 3.3.0dev documentation," 2025. Accessed: 2025-03-25.

[24] E. W. Deutsch, J. A. Vizcaíno, A. R. Jones, P.-A. Binz, H. Lam, J. Klein, W. Bittremieux, Y. Perez-Riverol, D. L. Tabb, M. Walzer, *et al.*, "Proteomics standards initiative at twenty years: current activities and future work," *Journal of proteome research*, vol. 22, no. 2, pp. 287–301, 2023.

[25] N. Hoffmann, J. Rein, T. Sachsenberg, J. Hartler, K. Haug, G. Mayer, O. Alka, S. Dayalan, J. T. Pearce, P. Rocca-Serra, *et al.*, "mztab-m: a data standard for sharing quantitative results in mass spectrometry metabolomics," *Analytical chemistry*, vol. 91, no. 5, pp. 3302–3310, 2019.

[26] H. N. Po and N. Senozan, "The henderson-hasselbalch equation: its history and limitations," *Journal of chemical education*, vol. 78, no. 11, p. 1499, 2001.

[27] M. Pridatchenko, I. Tarasova, V. Guryca, A. Kononikhin, C. Adams, D. Tolmachev, A. Y. Agapov, V. Evreinov, I. Popov, E. Nikolaev, *et al.*, "Use of models of biomacromolecule separation in amt database generation for shotgun proteomics," *Biochemistry (Moscow)*, vol. 74, pp. 1195–1202, 2009.

[28] A. A. Goloborodko, L. I. Levitsky, M. V. Ivanov, and M. V. Gorshkov, "Pyteomics—a python framework for exploratory data analysis and rapid software prototyping in proteomics," *Journal of The American Society for Mass Spectrometry*, vol. 24, no. 2, pp. 301–304, 2013.

[29] A. F. Jarnuczak and J. A. Vizcaíno, "Using the pride database and proteomexchange for submitting and accessing public proteomics datasets," *Current protocols in bioinformatics*, vol. 59, no. 1, pp. 13–31, 2017.

[30] N. F. de Jonge, H. Hecht, M. Strobel, M. Wang, J. J. van der Hooft, and F. Huber, "Reproducible ms/ms library cleaning pipeline in matchms," *Journal of Cheminformatics*, vol. 16, no. 1, p. 88, 2024.

Overview of Foundation Models in Remote Sensing

1st Teo Stojkovski
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
teo.stojkovski@students.finki.ukim.mk

2nd Evrosina Stojkoska
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
evrosina.stojkoska@students.finki.ukim.mk

3rd Marko Petrov
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
marko.petrov@finki.ukim.mk

4th Ema Pandilova
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
ema.pandilova@finki.ukim.mk

5th Vlatko Spasev
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
vlatko.spasev@finki.ukim.mk

6th Ivan Kitanovski
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
ivan.kitanovski@finki.ukim.mk

***Abstract*—Recent advances in artificial intelligence, particularly the rise of foundation models, are reshaping the landscape of remote sensing. Traditionally reliant on handcrafted features and task-specific models, remote sensing has faced challenges in scalability, adaptability, and generalization across diverse geospatial data sources. Foundation models—large-scale, pre-trained architectures developed using self-supervised learning—offer a transformative solution by enabling cross-task generalization, improved performance in object detection and segmentation, and reduced reliance on annotated datasets. This paper provides a comprehensive review of foundation models in the context of remote sensing, highlighting their architectures, training strategies, and applications across environmental monitoring, agriculture, disaster response, and urban planning. We also address the technical and practical challenges of integrating these models into heterogeneous geospatial workflows, including variations in data resolution, sensor modalities, and real-time processing demands.**

***Index Terms*—Remote Sensing Images, Object Detection, Deep Learning, Earth Observation**

I. Introduction

The rapid advancements in artificial intelligence and deep learning have significantly transformed the field of remote sensing, enabling the extraction of meaningful insights from vast amounts of satellite and aerial imagery [1]. As an essential technology for Earth observation, remote sensing plays a crucial role in various domains, including environmental monitoring, agriculture, disaster management, and urban planning. However, traditional remote sensing methods often rely on handcrafted features and task-specific models, which can be limited in scalability, generalization, and adaptability to diverse geospatial data sources.

The emergence of foundation models [2] has introduced a paradigm shift in remote sensing by leveraging large-scale pre-trained architectures capable of generalizing across multiple downstream tasks. These models, trained on extensive datasets using self-supervised learning techniques, offer significant improvements in object detection, image segmentation, and change detection by effectively capturing spatial and temporal patterns in geospatial data. Unlike conventional deep learning approaches that require extensive labeled datasets for each specific application, foundation models enable knowledge transfer [3] and zero-shot [4] learning, reducing the dependency on annotated data while enhancing model robustness.

Despite these advantages, integrating foundation models into remote sensing presents unique challenges. Remote sensing data varies widely in resolution, color bands, and sensor types, making it challenging to process using a single approach. To address this, model architectures must be designed to handle different types of inputs effectively, ensuring accurate analysis across diverse geospatial datasets [5]. Additionally, computational efficiency and the need for domain-specific adaptations remain key concerns, particularly in high-resolution satellite imagery applications where real-time processing is crucial.

This paper aims to offer a thorough overview of foundation models in remote sensing, exploring their architecture, training methodologies, and performance across various geospatial tasks. We examine their impact on traditional remote sensing workflows, assess the latest advancements in the field, and discuss existing challenges and potential future developments.

The paper is organized as follows: In the second section we provide a background in the topic of remote sensing, the data types present, the basic concepts of processing the data and challenges. The third section provides an overview of the foundation models in the context of remote sensing, explaining their details and characteristics. In the fourth section, we com-

pare and discuss the different models and their capabilities. The concluding remarks are provided in the final section.

II. Background

A. Remote sensing data modalities

Remote sensing was limited and time-consuming in the past, mainly due to its reliance on data acquired through aerial photography and ground-based analysis. With the refinement of sensor technology and satellite imagery, modern remote sensing has revolutionized how we view and understand the surface of the Earth. Today, vast amounts of data are being collected through an array of advanced sensors, enabling precise monitoring of environmental and man-made changes.

In order to capture the features of the Earth's surface, remote sensing uses a variety of data modalities (Figure 1), each having unique imaging methods, spectral ranges, and spatial resolution. These data types are essential for applications such as climate monitoring, disaster response, agriculture, and urban planning. In this section, we briefly analyze the following types of data modalities:

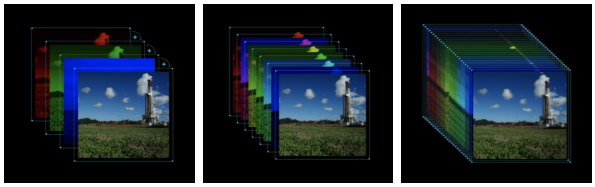
Optical RGB (**R**ed, **G**reen, and **B**lue) imagery [6] plays a fundamental role in remote sensing by capturing data in the visible spectrum and providing full-color images that closely resemble what the human eye sees. It is a useful tool for numerous applications, such as urban planning, disaster response, and object detection, due to its accessibility and easily interpreted output. RGB imagery is frequently obtained from multiple platforms, such as unmanned aerial vehicles (UAVs) and drones for regional observations, planes for large-scale object detection, and satellites for global remote sensing. It usually uses an 8-bit per channel system, where each channel records intensity values that define an image's color composition and brightness. These values range from 0 (black) to 255 (white). However, despite its technical simplicity, RGB imagery has limitations, particularly its sensitivity to lighting conditions and its restricted spectral range compared to multispectral and hyperspectral imaging.

Multispectral remote sensing [7] acquires data across the visible and infrared spectrum, with varying spatial and spectral resolutions. In comparison with hyperspectral imagery, multispectral images typically occur with lower spectral, but higher spatial resolution, typically ranging from 5 to 12 bands [8]. This makes multispectral data popular for long-term environmental research. These data are obtained through various platforms, including satellites such as Landsat, Sentinel-2, and MODIS.

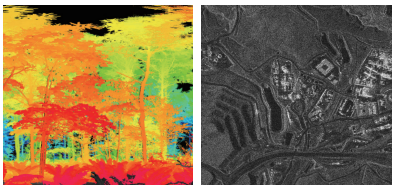
Hyperspectral remote sensing [8] captures data in a profoundly ranged specter, offering images with high spectral resolution. Compared to multispectral imagery, which provides data in a limited number of broad bands, hyperspectral data has much higher band number, ranging from hundreds to thousands. This spectral resolution allows the detection of information that is invisible to multispectral sensors, making hyperspectral data particularly valuable for complex data analysis in fields such as agriculture, forestry and environmental monitoring.

LiDAR (**L**ight **D**etection and **R**anging) technology [9] uses laser pulses for measuring distances, by capturing their reflection from objects in the environment. In remote sensing, this technology enables high-resolution 3D mapping and is widely used for capturing terrain data, water structures, man-made objects, and detecting deformations in structural engineering. LiDAR systems are deployed on terrestrial platforms, unmanned aerial vehicles (UAVs), aircraft, and satellites, supporting various applications such as urban planning, forestry, and object detection.

SAR (**S**ynthetic **A**perture **R**adar) [10] is a widely used data modality in remote sensing due to its ability to operate in all weather conditions and at any time of day. Unlike optical sensors, which rely on sunlight, SAR emits microwave signals and measures their returns, allowing it to penetrate certain surfaces. This makes SAR particularly valuable for applications such as environmental monitoring, disaster assessment, land cover classification, and infrastructure mapping.



(a) RGB (b) Multispectral (c) Hyperspectral



(d) LiDAR (e) SAR

Fig. 1. Remote sensing data modalities: (a) RGB, (b) Multispectral, (c) Hyperspectral, (d) LiDAR [11], (e) SAR. Image sourced from: https://www.critchlow.co.nz/resources/blog/next-generation-satellite-services-high-definition-imagery-daily-change-detection-multi-spectral-derived-data, https://www.cg.informatik.uni-siegen.de/de/exploration-synthetic-aperture-radar-sar-images

B. Remote sensing processing concepts

Data gathered through remote sensing must be processed to extract meaningful insights. Traditionally, this process relied on manual interpretation and task-specific models designed for specific applications. While traditional methods are straightforward, they often require extensive labeled datasets and significant data processing capabilities [12].

With the rise of deep learning, remote sensing has undergone a revolution, enabling the automation of data analysis with improved accuracy and efficiency.

Vision Transformers (**ViTs**) have proven to be a high-performance architectural framework for remote sensing image analysis and classification. ViTs process images by dividing them into small patches, allowing the model to focus on patterns and details within each section, leading to classification

accuracies of up to 98.49%. This architectural design makes ViTs highly effective at classifying complex remote sensing imagery [13].

C. Challenges in remote sensing

Due to the complexity of data in remote sensing, several challenges arise during image processing. Image classification relies on labeled samples, typically obtained through ground surveys or image photointerpretation. However, ground surveys offer accuracy with high cost, while photointerpretation is more affordable but has low precision. Variations in spectral signatures caused by environmental factors, seasonal changes, and sensor differences, complicate image segmentation. Although deep neural networks (DNNs) have enhanced segmentation, they are limited by the low volume of large labeled datasets. Object detection faces reduced detection accuracy due to the variation of object sizes, shapes, and spectral features, as well as the interference of environmental noise, clouds, and shadows. Traditional algorithms such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) often require voluminous labeled datasets, which challenges their ability to detect small or obscured objects [14].

III. Foundation models in Remote Sensing

A. What are Foundation models and why do we need them?

Foundation models (FMs) are large-scale, pre-trained models designed to serve as a foundation for various tasks across different field [15]. These models are trained on vast datasets using advanced architectures, allowing them to learn complex patterns and features. This pre-training makes them adaptable to specific applications with minimal additional training. Techniques like self-supervised learning (SSL) [16] and transformer-based [17] architectures have significantly improved their effectiveness in tasks such as image classification, object detection, and change detection.

One of the key advantages of FMs is their ability to learn from largely unlabeled data using SSL. This is especially beneficial in remote sensing, where acquiring labeled datasets is often challenging [18].

Several notable foundation models have been developed specifically for remote sensing, such as SatMAE which is a transformer-based model designed for temporal and multispectral satellite imagery [19].

Despite their advancements, FMs still face challenges, including the need for high-quality and diverse training data, significant computational resources, and effective adaptation to specific remote sensing applications [20]. Overcoming these limitations will be essential for the continued progress of foundation models in remote sensing.

B. Types of Models

Early models relied on convolutional neural networks (CNNs), such as ResNet [21], which improved image recognition and classification [22]. Vision transformers (ViTs) adapted this approach for image data by treating image patches as sequences of tokens, allowing models to recognize both local and global relationships [23]. This is particularly beneficial for tasks like semantic segmentation and change detection, which require understanding fine details in high-resolution satellite images.

Most foundation models [24] cannot perform zero-shot inference due to the lack of joint vision-language modeling. To address these challenges, researchers developed a vision-language foundation model for remote sensing, named RemoteCLIP. The goal was to learn strong visual features with rich semantics from satellite imagery while aligning these features with text embeddings [25].

Many of the current methods are inspired by self-supervised learning (SSL) techniques in computer vision, especially Masked Image Modeling (MIM) [26] [27] [28]. Recent models like SatMAE [19], Scale-MAE [29], ViTAE [30], Billion-scale MAE [31], RingMo [32], and GFM [33] have successfully leveraged MIM with large Vision Transformers (ViTs) trained on vast satellite imagery datasets, yielding promising outcomes.

C. Foundation Models in Remote Sensing

Remote sensing (RS) is the process of gathering information about objects or areas from a distance, usually through satellite or airborne sensors [34].

Large Foundation Models (FMs) have demonstrated remarkable generalization and zero-shot learning abilities in multiple domains [35]. These models learn generalized representations of Earth's surface features and can be fine-tuned for a wide range of downstream tasks, such as land cover classification, object detection, change detection, and environmental monitoring. By leveraging transfer learning, foundation models reduce the need for extensive labeled data in domain-specific applications and significantly accelerate development cycles. Their ability to generalize across sensors, regions, and modalities (e.g., optical, SAR, hyperspectral) makes them especially powerful for global-scale analysis and applications in agriculture, disaster response, climate monitoring, and more.

Several notable foundation models have been developed specifically for remote sensing [25], such as SatMAE, Scale-MAE, ViTAE, Billion-scale MAE, RingMo, GFM, RemoteCLIP, SimCLR, MAE, SAM, and BERT. These models are often trained on large-scale, multi-sensor datasets, enabling them to extract rich, transferable representations from diverse imagery sources. Some, like RemoteCLIP and SimCLR, focus on contrastive learning approaches, while others, like MAE and SatMAE, use masked autoencoding to learn robust visual features. Their architectures are increasingly incorporating modality-specific adaptations to better handle unique remote sensing challenges, such as varying spatial resolutions, spectral bands, and geospatial context

IV. Discussion

The emergence of foundation models has led to significant advancements in remote sensing. Using deep learning architectures, these models have demonstrated remarkable improvements in the field, enhancing the accuracy and efficiency of

TABLE I
COMPARISON OF REMOTE SENSING MODELS BASED ON REFERENCES, PARAMETERS, PRETRAINING DATASET SIZE, MODALITIES, AND AVAILABILITY OF WEIGHTS AND ARCHITECTURE

Model	References	Parameters	Pretraining Dataset Size	Modalities	Weights Available	Architecture Available
SeCo	313	23.5M	1,000,000	Multispectral	✓	✓
GeoKR	275	138M	1,431,950	RGB, Multispectral	×	✓
RVSA	269	100M	1,000,848	RGB, Multispectral, SAR	✓	✓
SatMAE	265	307M	1,047,691	RGB, Multispectral, Temporal	✓	✓
RingMo	244	10B	2,000,000	RGB, SAR	✓	✓
RSP	219	25.8M	1,000,848	Multispectral, Hyperspectral	✓	✓
ScaleMAE	172	322.9M	1,047,691	RGB, Multispectral	×	✓
SkySense	112	2.06B	21,500,000	RGB, Multispectral, SAR	✓	×
Prithvi	98	100M	4,200,000	Multispectral	✓	✓
GFM	78	100M	600,000	RGB, Multispectral, SAR	✓	✓

tasks such as image classification, segmentation, and object detection. This section will examine the findings of recent studies, compare performance evaluations, and analyze their overall impact on remote sensing. Since, there are numerous foundation models, we decided to focus on the top 10 models by the number of references of their respective publications.The number of references was determined based on citation counts retrieved from Google Scholar, ensuring a consistent and objective comparison across models. Table I provides a summary of the key attributes of these foundation models, including references, parameters, dataset sizes, and modalities, highlighting the diversity of models used across various remote sensing tasks.

A. Performance of Foundation Models Across Tasks

The performance metrics presented here are directly sourced from original studies, ensuring accuracy and consistency in evaluating these foundation models on benchmark datasets. This provides key insights into their effectiveness and limitations in various tasks.

In **image classification**, foundation models have demonstrated strong performance on the BigEarthNet dataset [36]. Models like msGFM [37] and Skysense [38] achieved the highest mAP of 92.90% and 92.09%, respectively. Their superior performance makes them well suited for high-precision applications. Other models, such as DeCUR [39] (89.70%) and DINO-MC [40] (88.75%), have also shown strong classification capabilities. The variations in performance highlight the importance of model specialization, with some models excelling in specific environmental conditions [41].

Semantic segmentation. Evaluations on ISPRS Postdam dataset show that SkySense [38] achieved the highest mF1 score of 93.99%, while CMID [42] attained the highest mIoU score of 87.04%, making them highly effective in capturing detailed spatial patterns. Models like BFM, Cross-scale MAE, and UPetu, also demonstrate competitive segmentation abilities, highlighting the significant variations in model performance across segmentation tasks [41].

For **object detection**, foundation models were evaluated on the DOTA dataset, where RVSA [30] achieved the highest mAP of 81.24%, followed by SMLFR [43] (79.33%) and RSP [44] (77.72%). On the DIOR and DIOR-R datasets, MTP [45] and SkySense [38] demonstrated strong performance, with an AP50 of 78% and a mAP of 78.73%, respectively. These results highlight the ability of these models to accurately detect objects across various remote sensing applications [41].

In **change detection** tasks, models such as SkySense [38] and GFM [33] achieved the highest F1 scores of 60.06% and 59.82% on the OSCD dataset. Meanwhile, performance was significantly higher on the LEVIR-CD dataset, with MTP [45] attaining an F1 score of 92.67%, closely followed by SkySense [38] (92.58%) [41].

B. Influence of pre-training methods

The rapid advancement of pre-training techniques has greatly expanded the capabilities of foundation models in remote sensing, allowing them to extract complex features from vast amounts of data. By leveraging self-supervised learning techniques, such as contrastive learning (CL) and masked autoencoders (MAE), foundation models have outperformed traditional learning methods in every aspect.

Skysense, which implements a multi-granularity contrastive learning, achieves approximately 3.6% higher accuracy in scene classification and object detection [38] compared to other conventional models. Similarly, Seco, utilizing seasonal contrast learning, improves land-cover classification by up to 7% over ImageNet pre-trained models [46]. Models such as SatMAE and ScaleMae, which utilize masked autoencoding, have demonstrated advancements in multi-temporal and multispectral data processing. SatMae [19] achieved an improvement of up to 14% in land cover classification, while ScaleMae [29] showed an increase in mIoU by 1.7% in segmentation precision.

Among these foundation models, RingMo stands out as the largest and most complex, with 10 billion parameters. Meanwhile, SkySense was trained on the largest dataset, or a combination of multiple datasets, consisting of over 21.5 million images, making it highly effective for multi-modal remote sensing tasks. Overall, these results underscore the importance of pre-training methods in maximizing the performance of foundation models, making them more reliable and efficient in handling complex data [41].

C. Practical Implications and Applications

Foundation models transform the field of remote sensing by enhancing traditional applications like multi-spectral and time-

series data analysis. These models offer higher performance with less reliance on huge labeled data. In this section, we will briefly discuss some of the practical implications of foundation models across different areas:

- **Environmental Monitoring**. Foundation models play a fundamental role in monitoring deforestation, desertification, and pollution. Models like GASSL [47] and SatMAE [48] use multi-spectral and temporal data to detect environmental degradation early, reducing the response time and minimizing the risk of larger disasters.
- **Agriculture and Forestry**. Models like EarthPT [49] and GeCo [50] provide valuable information on crop health, yield prediction, and land use management. RSP [44], for instance, enables precise monitoring of crop conditions, allowing for the early detection of crop stress signs. These models are also highly effective in forestry management, providing detailed maps of forest cover, biomass estimation, and deforestation monitoring.
- **Archaeology**. Models such as GeoKR [51] and RingMo [32] employ high-resolution satellite images and multispectral data to detect archaeological features that are difficult to identify with the naked eye. Meanwhile, models like MATTER [52] analyze texture and material compositions, offering insights into various surface characteristics.
- **Urban Planning and Development**. Foundation models are essential for monitoring urban expansion, infrastructure growth, and land use changes. CMID [42] and SkySense [38] are models that allow advanced urban planning, providing insights on the impact of urbanization on natural habitats and help city planners optimize land use decisions.
- **Disaster Management**. Remote sensing models such as OFA-Net [53], DOFA [54] and Prithvi [55] provide crucial real-time data for detecting affected areas, minimizing response time, and supporting emergency responders in resource allocation and evacuation planning. Furthermore, these models also play a vital role in post-disaster recovery by assessing damages and monitoring environmental changes.

V. Conclusion

In this paper, we explored various foundation models, highlighting their role in remote sensing applications. Most of these models utilize RGB and multispectral data to address tasks such as object detection, land cover classification, and change detection. Their ability to process diverse data modalities makes them valuable for applications in agriculture, urban planning, environmental monitoring, and disaster management. Additionally, this paper emphasizes the importance of pretraining these models on large-scale datasets to enhance their generalization and performance across different remote sensing tasks. By understanding their capabilities and limitations, we can effectively use these models to enhance remote sensing analysis and decision-making.

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] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE geoscience and remote sensing magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[2] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundation models defining a new era in vision: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[4] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

[5] H. Wang, E. Skau, H. Krim, and G. Cervone, "Fusing heterogeneous data: A case for remote sensing and social media," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 6956–6968, 2018.

[6] A. Kior, L. Yudina, Y. Zolin, V. Sukhov, and E. Sukhova, "Rgb imaging as a tool for remote sensing of characteristics of terrestrial plants: A review," *Plants*, vol. 13, no. 9, p. 1262, 2024.

[7] Y. Zhao, X. Zhang, W. Feng, and J. Xu, "Deep learning classification by resnet-18 based on the real spectral dataset from multispectral remote sensing images," *Remote Sensing*, vol. 14, no. 19, p. 4883, 2022.

[8] T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, and J. J. Sousa, "Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry," *Remote sensing*, vol. 9, no. 11, p. 1110, 2017.

[9] Z. Wang and M. Menenti, "Challenges and opportunities in lidar remote sensing," *Frontiers in Remote Sensing*, vol. 2, p. 641723, 2021.

[10] C.-a. Liu, Z.-X. Chen, S. Yun, J.-s. Chen, T. Hasi, and H.-z. Pan, "Research advances of sar remote sensing for agriculture applications: A review," *Journal of integrative agriculture*, vol. 18, no. 3, pp. 506–525, 2019.

[11] K. Omasa, F. Hosoi, and A. Konishi, "3d lidar imaging for detecting and understanding plant responses and canopy structure," *Journal of experimental botany*, vol. 58, no. 4, pp. 881–898, 2007.

[12] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, P. VanValkenburgh, S. A. Wernke, and Y. Huo, "Ai foundation models in remote sensing: A survey," *arXiv preprint arXiv:2408.03464*, 2024.

[13] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, p. 516, 2021.

[14] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016.

[15] L. Jiao, Z. Huang, X. Lu, X. Liu, Y. Yang, J. Zhao, J. Zhang, B. Hou, S. Yang, F. Liu *et al.*, "Brain-inspired remote sensing foundation models and open problems: A comprehensive survey," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 10 084–10 120, 2023.

[16] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He *et al.*, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *International Journal of Machine Learning and Cybernetics*, pp. 1–65, 2024.

[19] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197–211, 2022.

[20] M. Noman, M. Naseer, H. Cholakkal, R. M. Anwer, S. Khan, and F. S. Khan, "Rethinking transformers pre-training for multi-spectral satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 811–27 819.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Y. Ma, S. Chen, S. Ermon, and D. B. Lobell, "Transfer learning in environmental remote sensing," *Remote Sensing of Environment*, vol. 301, p. 113924, 2024.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[25] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[27] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[28] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9653–9663.

[29] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4088–4099.

[30] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2022.

[31] K. Cha, J. Seo, and T. Lee, "A billion-scale foundation model for remote sensing images," *arXiv preprint arXiv:2304.05215*, 2023.

[32] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2022.

[33] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Towards geospatial foundation models via continual pretraining," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 806–16 816.

[34] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, S. A. Wernke, Y. Huo *et al.*, "Vision foundation models in remote sensing: A survey," *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[35] A. Xiao, W. Xuan, J. Wang, J. Huang, D. Tao, S. Lu, and N. Yokoya, "Foundation models for remote sensing and earth observation: A survey," *arXiv preprint arXiv:2410.16602*, 2024.

[36] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*. IEEE, 2019, pp. 5901–5904.

[37] B. Han, S. Zhang, X. Shi, and M. Reichstein, "Bridging remote sensors with multisensor geospatial foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 852–27 862.

[38] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 672–27 683.

[39] Y. Wang, C. M. Albrecht, N. A. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, "Decur: decoupling common & unique representations for multimodal self-supervision," 2023.

[40] X. Wanyan, S. Seneviratne, S. Shen, and M. Kirley, "Extending global-local view alignment for self-supervised learning with remote sensing imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2443–2453.

[41] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, P. VanValkenburgh, S. A. Wernke, and Y. Huo, "Vision foundation models in remote sensing: A survey," *arXiv preprint arXiv:2408.03464*, 2024.

[42] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, "Cmid: A unified self-supervised learning framework for remote sensing image understanding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[43] Z. Dong, Y. Gu, and T. Liu, "Generative convnet foundation model with sparse modeling and low-frequency reconstruction for remote sensing image interpretation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[44] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2022.

[45] D. Wang, J. Zhang, M. Xu, L. Liu, D. Wang, E. Gao, C. Han, H. Guo, B. Du, D. Tao *et al.*, "Mtp: Advancing remote sensing foundation model via multi-task pretraining. arxiv," *arXiv preprint arXiv:2403.13430*, 2024.

[46] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.

[47] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 181–10 190.

[48] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[49] A. Stewart, N. Lehmann, I. Corley, Y. Wang, Y.-C. Chang, N. A. Ait Ali Braham, S. Sehgal, C. Robinson, and A. Banerjee, "Ssl4eo-l: Datasets and foundation models for landsat imagery," *Advances in Neural Information Processing Systems*, vol. 36, pp. 59 787–59 807, 2023.

[50] W. Li, K. Chen, and Z. Shi, "Geographical supervision correction for remote sensing representation learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[51] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[52] P. Akiva, M. Purri, and M. Leotta, "Self-supervised material and texture representation learning for remote sensing tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8203–8215.

[53] Z. Xiong, Y. Wang, F. Zhang, and X. X. Zhu, "One for all: Toward unified foundation models for earth vision," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 2734–2738.

[54] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. L. Saux, G. Camps-Valls, and X. X. Zhu, "Neural plasticity-inspired multimodal foundation model for earth observation," *arXiv preprint arXiv:2403.15356*, 2024.

[55] J. Jakubik, S. Roy, C. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards *et al.*, "Foundation models for generalist geospatial artificial intelligence," *arXiv preprint arXiv:2310.18660*, 2023.

Accelerating IoV Intrusion Detection: Benchmarking GPU-Accelerated vs CPU-Based ML Libraries

1st Furkan Çolhak
CCIP, Center for Cyber Security and Critical Infrastructure Protection, Kadir Has University
Istanbul, Turkey
furkancolhak@stu.khas.edu.tr

2nd Hasan Coşkun
CCIP, Center for Cyber Security and Critical Infrastructure Protection, Kadir Has University
Istanbul, Turkey
hasan.coskun@stu.khas.edu.tr

3rd Tsafac Nkombong Regine Cyrille
CyberMACS, Kadir Has University
Istanbul, Turkey
tsafacnkombong.cyrille@stu.khas.edu.tr

4th Tedi Hoxa
CyberMACS, Kadir Has University
Istanbul, Turkey
tedihoxa@stu.khas.edu.tr

5th Mert İlhan Ecevit
CCIP, Center for Cyber Security and Critical Infrastructure Protection, Kadir Has University
Istanbul, Turkey
mertilhan.ecevit@khas.edu.tr

6th Mehmet Nafiz Aydın
CyberMACS, Kadir Has University
Istanbul, Turkey
mehmet.aydin@khas.edu.tr

***Abstract*—The Internet of Vehicles (IoV) may face challenging cybersecurity attacks that may require sophisticated intrusion detection systems, necessitating a rapid development and response system. This research investigates the performance advantages of GPU-accelerated libraries (cuML) compared to traditional CPU-based implementations (scikit-learn), focusing on the speed and efficiency required for machine learning models used in IoV threat detection environments. The comprehensive evaluations conducted employ four machine learning approaches (Random Forest, KNN, Logistic Regression, XGBoost) across three distinct IoV security datasets (OTIDS, GIDS, CICIoV2024). Our findings demonstrate that GPU-accelerated implementations dramatically improved computational efficiency, with training times reduced by a factor of up to 159 and prediction speeds accelerated by up to 95 times compared to traditional CPU processing, all while preserving detection accuracy. This remarkable performance breakthrough empowers researchers and security specialists to harness GPU acceleration for creating faster, more effective threat detection systems that meet the urgent real-time security demands of today's connected vehicle networks.**

***Index Terms*—Internet of Vehicles (IoV), Intrusion Detection System (IDS), GPU Acceleration, Machine Learning, cuML, scikit-learn, Training Time, Prediction Time, Computational Efficiency**

I. Introduction

In recent years, the Internet of Things (IoT) has gained widespread adoption, connecting devices to improve efficiency in sectors such as transportation, healthcare and energy [1]. A key extension of IoT is the Internet of Vehicles (IoV), which integrates vehicles into smart automotive environments to support real-time monitoring, predictive maintenance, and autonomous driving [2]. Despite the fact that IoV brings significant benefits, it also faces some important cybersecurity challenges [3].

Since IoV vehicles can directly affect the safety of life and property, it is important for IoV vehicles to implement strong security measures, with Intrusion Detection Systems (IDS) being essential for detecting and mitigating cyber threats. A notable example of these risks occurred in 2015, when security researchers demonstrated a critical vulnerability (CVE-2015-5611) in Fiat Chrysler's Uconnect system, affecting vehicles manufactured between 2013 and 2015. The researchers remotely compromised a 2014 Jeep Cherokee through its entertainment system, gaining access to critical vehicle controls, including steering, brakes, and transmission [4]. Exploitation of this vulnerability may allow an unauthorized user to take remote control of an affected vehicle, but the attack requires access. This incident led to the recall of 1.4 million vehicles and highlighted the severe consequences of insufficient security measures in connected vehicles. More recently, a Volkswagen data breach exposed sensitive information about 800,000 owners of electric vehicles, including detailed GPS locations and usage patterns [5]. These incidents emphasize the crucial importance of adopting robust and innovative IoV cybersecurity measures and effective intrusion detection systems (IDS).

The increasing sophistication of IoV attacks, combined with the massive scale and complexity of vehicle networks, creates unique challenges for intrusion detection systems, particularly in the context of real-time threat detection. Although traditional IDS approaches have proven effective in conventional networks, vehicular networks' real-time requirements and complex nature demand more efficient computational solutions. Machine learning-based IDS approaches have shown promising results, with techniques such as Extreme Gradient

TABLE I
OVERVIEW OF INTRUSION DETECTION STUDIES IN VEHICLE SYSTEMS (CONDENSED MODELS)

Author(s)	Study Summary (incl. Comp. Time/Efficiency Focus)	ML Models	Libraries	Dataset
Purohit and Govindarasu (2022) [6]	The study optimizes training and execution time by using efficient tree-based algorithms like Decision Tree, Random Forest, and XGBoost, combined with a rule-based system, achieving high accuracy and low latency for real-time CAN bus intrusion detection.	DT, RF, XGB	Keras, Pandas	OTIDS
Saber and Mazri (2023) [7]	The study emphasizes the importance of efficient feature selection and classifier optimization to enhance detection accuracy and reduce computational overhead.	KNN, DT, NB, SVM, LR, RF	NumPy, SciPy, Scikit-learn	GIDS
Liu and Fan (2023) [8]	The study employs Recursive Feature Elimination (RFE) to reduce data dimensions, improving detection efficiency. Decision Trees and Multi-Layer Perceptron (MLP) are used for classification and optimizing execution speed.	DT, MLP	*N*ot explicitly mentioned	CICIDS17, GIDS
T.-T.-H. Le et al. (2023) [9]	Proposes XGBoost with SHAP explanations for IDS on imbalanced CAN data. Focuses on binary/multiclass accuracy and interpretability. Computational time/efficiency is not the primary focus, but efficiency is inherent in choosing XGBoost over some DL methods.	XGB	Scikit-learn (implied)	GIDS
Limouchi & Chan (2023) [10]	Optimized ML (LightGBM, ET, XGBoost) for IoV IDS. Addresses imbalance (SMOTE), uses Correlation-based Feature Selection (FS), and Bayesian Optimization (BO) for hyperparameters. Focuses on performance enhancement and reducing computational cost via FS.	LGBM, ET, XGB	Pandas, Matplotlib, Seaborn, Scikit-learn	CICDDoS2019
Kumar & Singh (2023) [11]	Uses tree-based ML (DT, RF, ET, XGBoost) and Stacking ensemble for AV IDS. Addresses imbalance (SMOTE) and uses feature selection. Aims for high detection rates at low processing costs, highlighting computational efficiency as a consideration.	DT, RF, ET, XGB	Not explicitly mentioned	CICIDS2017, GIDS
Gou et al.(2023) [12]	Proposes an Adaptive Tree-Based Ensemble Network (ATBEN) with stacked ML models for multiclass IDS in IoV. Addresses imbalance (SMOTE + RUS) and uses ML-based feature selection (FS). Aims for efficient multiclass classification, mentions reducing computational overhead.	ATBEN (RF, ET, XGB, LGBM etc.)	PyTorch	CICIDS2017, GIDS
Kaushik et al. (2024) [13]	Optimizes training and execution time using feature selection with Mutual Information and Correlation, achieving significant reductions in computational time across various classifiers.	DT, LR, LDA, NB	Scikit-learn (implied)	CICIDS18
Korium et al. (2024) [14]	**U**tilizes ML models with Z-score normalization and regression-based feature selection to enhance model efficiency, achieving high accuracy and optimized execution and detection times.	RF, XGB, CatBoost, LGBM	Scikit-learn	CICIDS17, CSEIDS18, CICDDoS19

Boosting and tree-based models demonstrating high detection accuracy [15], [16]. However, the computational demands of these ML models present a significant challenge in IoV environments, where real-time detection is crucial [13]. The choice between GPU-accelerated and CPU-based implementations can significantly impact both detection accuracy and response time, yet there is limited research comparing their performance in IoV contexts.

This paper addresses this challenge by investigating the performance of machine learning models in IoV environments, focusing on the computational efficiency between GPU-accelerated libraries (cuML) and CPU-based implementations (scikit-learn). Using three distinct IoV datasets: CAN-intrusion dataset (OTIDS), Car-hacking dataset (GIDS), and CICIoV2024. This work evaluates the models using key metrics, including accuracy, F1 score, training time, prediction time and speedup metrics.

Our research has made contributions that can be outlined as follows:

- This study systematically compares GPU-accelerated (cuML) versus CPU-based (Scikit-learn) machine learning for IoV tasks, analyzing both performance (accuracy, F1 score) and computational efficiency (training/prediction times, speedup factors).
- The evaluation uses three diverse IoV datasets (OTIDS, GIDS, and CICIoV2024) to comprehensively and realistically assess model suitability in different scenarios.

II. LITERATURE REVIEW

A. Research Methodology

In conducting a comprehensive review of the literature on Intrusion Detection Systems (IDS) within the Internet of Vehicles (IoV) framework, this work utilized multiple databases to ensure a broad coverage of related studies.

These databases included IEEE Xplore, Scopus, Web of Science, and ACM Digital Library. The queries were constructed around key elements of our research focus, covering IoV concepts, security and IDS, machine learning techniques, data challenges, and specific datasets. Additionally, the search was restricted to journal and conference papers published after 2022.

The following queries were employed:

- **Query 1:** `"Internet of Vehicles" OR "IoV" OR "Vehicular Network" OR "Connected Vehicles"`
- **Query 2:** `"Intrusion Detection System" OR "IDS" OR "Cybersecurity" OR "Attack Detection"`
- **Query 3:** `"Machine Learning" OR "ML"`
- **Query 4:** `"Training Time" OR "Execution Time" OR "Computational Efficiency"`

The initial search produced many results: 16,218 publications across all databases for the broadest query set. Applying the subsequent filters and combining queries gradually nar-

TABLE II
IoV IDS Dataset Properties

Dataset	Year	Sample Count	Class	Count	%
CAN-intrusion-dataset (OTIDS) [17]	2017	4,613,439	Benign	2,369,398	51.35
			DoS	656,579	14.23
			Fuzzy	591,990	12.83
			Impersonation	995,472	21.57
Car-Hacking Dataset (GIDS) [18]	2018	14,427,180	Benign	2,056,938	6.85
			RPM	4,621,702	32.03
			DoS	3,665,771	25.40
			Fuzzy	2,056,938	14.25
			Gear	3,093,898	21.44
CICIoV2024 [3]	2024	1,408,219	Benign	1,223,737	86.89
			DoS	74,663	5.30
			RPM	54,900	3.89
			Speed	24,951	1.77
			Steering Wheel	19,977	1.41
			Gas	9,991	0.70

rowed this number. Combining Query 1 and Query 2 reduced the results to 1,206 studies. Further refinement by adding Query 3 led to 348 publications. The most specific search, incorporating Queries 1, 2, 3, and 4, resulted in only 24 studies. After a thorough review to eliminate duplicates and exclude studies not focused on IoV or IDS, only seven studies were directly relevant for detailed analysis. These seven papers were carefully selected based on their focus on IoV security, the use of machine learning algorithms and libraries, and their discussion of computational aspects such as training and execution time.

B. Review of the Studies

Increased vehicle connectivity within the Internet of Vehicles (IoV) expands the attack surface, necessitating robust Intrusion Detection Systems (IDS). Machine Learning (ML) is widely explored for effective IDS development. Table I overviews common ML approaches for detecting malicious activities in Controller Area Network (CAN) bus communications, where studies often address computational efficiency, including training and execution times.

A review of contemporary studies reveals a diverse range of ML approaches. Tree-based algorithms, such as Decision Trees (DT), Random Forest (RF), XGBoost (XGB), LightGBM (LGBM), and Extra Trees (ET), are frequently employed due to their efficiency and interpretability, as demonstrated by Purohit and Govindarasu (2022), Limouchi & Chan (2023), Kumar & Singh (2023), Korium et al. (2024), and Gou et al. (2023). Other classic classifiers like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), and Linear Discriminant Analysis (LDA) also find application (Saber and Mazri, 2023; Alshathri et al., 2024; Kaushik et al., 2024). More complex architectures, including Multi-Layer Perceptrons (MLP) (Liu and Fan, 2023), Graph Neural Networks (GNN) (He et al., 2023), and sophisticated ensemble methods like Stacking (Kumar & Singh, 2023) and adaptive tree-based networks (ATBEN) (Gou et al., 2023), are also being investigated. Interpretability is sometimes addressed using methods like SHAP (T.-T.-H. Le et al., 2023).

Computational efficiency is a critical consideration for real-time vehicular IDS. Several studies explicitly aim to optimize training and execution times. Purohit and Govindarasu (2022) focused on low latency using efficient tree models. Feature selection is frequently cited as a means to reduce computational load (Limouchi & Chan, 2023; Kaushik et al., 2024; Liu and Fan, 2023). Korium et al. (2024) specifically targeted optimized execution and detection times, while Kumar & Singh (2023) aimed for high detection rates at low processing costs. These optimizations are typically performed using standard CPU-based libraries, predominantly Scikit-learn, alongside libraries like Pandas and NumPy and specific ML framework libraries like Keras or PyTorch.

While these studies advance ML-based IDS for IoV, focusing on accuracy, imbalance, and feature dimensionality often within CPU constraints, there remains a need to explore alternative computational paradigms for further efficiency gains. The potential benefits of GPU acceleration, leveraging libraries specifically designed for accelerating data science workflows like cuML, have not been systematically compared against traditional CPU-based approaches (like Scikit-learn) for these specific IoV intrusion detection tasks.

Therefore, this study aims to bridge this gap by systematically comparing the performance (accuracy, F1 score) and computational efficiency (training time, prediction time, speedup factors) of GPU-accelerated (cuML) versus CPU-based (Scikit-learn) ML implementations. By utilizing diverse IoV datasets (OTIDS, GIDS, and CICIoV2024), this work provides a comprehensive assessment of the practical benefits and trade-offs of GPU acceleration for enhancing real-time intrusion detection capabilities in vehicular networks.

III. Datasets

The scarcity of IoV-based attack datasets masquerades as a primary challenge in security research. Due to this scarcity, researchers frequently depend on IoT-based datasets

like CICIDS2017 [19] and CICDDoS2019 [20], which contain traffic data for attack types such as brute force, DoS, and botnets. Whereas the datasets offer crucial insights for establishing baseline methodologies and evaluation frameworks, they break down address IoV-specific challenges, including vehicular network protocols and domain-specific vulnerabilities. To get meaningful and applicable results, working with domain-specific datasets is essential. This study focuses on IoV-specific datasets, ensuring intrusion detection research stays relevant and effective in addressing real-world security challenges [3].

Among the three datasets which are analyzed, stand out for their suitability: OTIDS [17], GIDS [18], and CICIoV2024 [3]. These datasets include real CAN data from various vehicles featuring benign samples and different types of attacks, such as DoS, fuzzing, spoofing, and impersonation, to evaluate intrusion detection systems and security mechanisms for vehicular networks. Further details about these datasets are presented in Table II.

A. Data Preprocessing

Data representations can vary significantly across datasets. In this research, the focus is on decimal data representation. For datasets that do not support decimal representation natively, conversion is straightforward. Initially, the datasets in this study were provided as separate files. All datasets were merged into one CSV file to simplify data processing and analysis. Notable attributes were retained, such as timestamps, IDs, DLC (Data Length Code), and data bytes (payloads). Also, the data bytes were split into individual bits and converted from hexadecimal to decimal format for better usability.

Some anomalies were identified in the OTIDS and GIDS datasets in the preprocessing phase. Irrelevant rows were removed, and missing values (NaN) were replaced with -1 to indicate missing data in a manner compatible with machine learning algorithms. Subsequently, timestamps and IDs were excluded from the datasets, as they were collected in a simulated environment and could inadvertently provide insights to machine learning models. Additionally, the DLC (Data Length Code) column was removed.

The final datasets contain only bit columns (DATA_0 to DATA_7) and the corresponding label. Data distributions are presented in Table II. Finally, object labels were numerically encoded using LabelEncoder, and features were standardized using StandardScaler for model training.

IV. Model Selection and Implementation

A. Machine Learning Algorithms

The following machine learning models were selected based on their effectiveness in intrusion detection tasks and relevance to IoV datasets.

Logistic Regression (LR) is a basic statistical model for classification, estimating probabilities to assign input data to categories. This study adapted it for multiclass classification using the OnevsOneClassifier, making it proper for detecting IoV attack types. While basic and efficient for smaller datasets, its performance degrades with increasing data size because of its linear nature, making it less competitive in complex IoV scenarios [21].

K-Nearest Neighbor (KNN) classifies data points based on their similarity to labeled neighbors, making it a simple yet flexible choice for multiclass problems. Although effective for smaller datasets, KNN deals with large datasets as the computational cost significantly increases the number of data points. In real-time IoV applications, this results in less scalability and slower forecasts [21].

Random Forest (RF) is an ensemble method that merges multiple decision trees, offering robustness to noise and adaptability to complex datasets. Its ability to manage high-dimensional data makes it ideal for IoV systems. On the other hand, its training times increase with data size and the number of trees, which can challenge its practicality in time-sensitive scenarios. [21].

Extreme Gradient Boosting (XGBoost) is a powerful enhancement of gradient boosting that sequentially builds decision trees to correct prior errors. Its ability to handle missing values and deliver high performance on structured datasets makes it particularly valuable for IoV systems. Despite these strengths, its computational complexity and reliance on careful hyperparameter tuning pose challenges for large datasets or limited computational resources [22].

B. Computational Libraries

This section provides an overview of the libraries used in this study. Pandas, NumPy, and scikit-learn were used for data preprocessing, numerical operations, and model evaluation.

scikit-learn (CPU-based): A widely used library for machine learning, scikit-learn includes algorithms like Logistic Regression, KNN, and Random Forest. It offers a wide range of parameters for tuning models, making it suitable for smaller datasets. However, it is optimized for CPU-based, single-threaded computation.

cuML (GPU-based): Part of the RAPIDS suite, cuML leverages NVIDIA CUDA for GPU acceleration, speeding up training and prediction for models like Random Forest, KNN, and Gradient Boosted Trees. It is highly efficient for large datasets, significantly reducing training time compared to CPU-based implementations, and supports hyperparameter tuning for enhanced model performance. However, it has fewer tuning parameters than scikit-learn, which may limit flexibility in certain use cases.

XGBoost is a separate, highly optimized library for gradient boosting; while it operates on the CPU by default, it also supports GPU acceleration through the device parameter or gpu_hist tree method.

V. Results and Discussion

A. Evaluation Metrics

Model evaluation centers on **Accuracy** and **F1-score** for performance assessment, where the F1-score's balance of precision/recall helps compare libraries robustly. Computational efficiency is measured by **training** and **prediction times**

TABLE III
MODEL PERFORMANCE COMPARISON ACROSS DIFFERENT DATASETS (INLINE SPEEDUP)

Dataset	Model	Library	Accuracy (%)	F1-Score (%)	Train Time (s)	Pred. Time (s)
OTIDS	RFC	Scikit-learn	82.37	82.17	279.28	5.49
		cuML	77.23	77.61	8.00 (↓ 34.9x)	0.28 (↓ 19.6x)
	kNN	Scikit-learn	81.76	81.57	11.47	48.23
		cuML	81.75	81.56	0.11 (↓ 104.3x)	1.43 (↓ 33.7x)
	LR	Scikit-learn	54.12	40.23	7.72	0.08
		cuML	54.12	40.24	1.15 (↓ 6.7x)	0.04 (↓ 2x)
	XGB	Scikit-learn	79.60	79.07	140.08	0.71
		cuML	79.60	79.07	26.34 (↓ 5.3x)	0.22 (↓ 3.2x)
GIDS	RFC	Scikit-learn	60.44	58.85	1603.99	46.19
		cuML	57.20	55.28	36.01 (↓ 44.5x)	0.48 (↓ 96.0x)
	kNN	Scikit-learn	50.77	51.49	44.56	27793.30
		cuML	49.82	50.18	0.28 (↓ 159.1x)	494.46 (↓ 56.2x)
	LR	Scikit-learn	38.09	33.68	25.58	0.27
		cuML	37.96	33.56	1.81 (↓ 14.1x)	0.07 (↓ 3.9x)
	XGB	Scikit-learn	60.39	58.61	214.42	2.14
		cuML	60.43	58.66	24.97 (↓ 8.6x)	0.51 (↓ 4.2x)
CICIoV2024	RFC	Scikit-learn	99.64	99.63	49.45	0.97
		cuML	99.29	99.24	4.33 (↓ 11.4x)	0.08 (↓ 12.1x)
	kNN	Scikit-learn	99.64	99.65	2.25	382.46
		cuML	99.64	99.63	0.04 (↓ 56.3x)	6.59 (↓ 58.0x)
	LR	Scikit-learn	87.47	84.14	19.87	0.02
		cuML	87.48	84.15	6.31 (↓ 3.1x)	0.02 (1.0x)
	XGB	Scikit-learn	99.64	99.65	134.17	0.12
		cuML	99.64	99.65	14.72 (↓ 9.1x)	0.11 (1.1x)

Note: Values in parentheses like (↓ Nx) next to cuML times indicate its speedup factor compared to Scikit-learn for that operation. The ↓ symbol denotes faster execution (less time).

(seconds), and the **speedup** is calculated as scikit-learn time / cuML time to quantify the acceleration provided by cuML.

B. Experimental Setup

Tests were conducted on a system with an Intel Core i9-10900X CPU (3.70 GHz) and an NVIDIA RTX A4000 GPU (16 GB VRAM, CUDA enabled). Datasets were stratified and randomly split into 80% training and 20% testing sets. Parameters are kept consistent across all models for fair comparison.

C. Performance Evaluation

This study evaluates the performance of Random Forest Classifier (RFC), k-Nearest Neighbors (kNN), Logistic Regression (LR), and eXtreme Gradient Boosting (XGB) models, comparing Scikit-learn (CPU) and cuML (GPU) implementations across the OTIDS, GIDS, and CICIoV2024 datasets.

As detailed in Table III, the primary advantage of cuML was a consistent and significant acceleration in training and prediction times across all datasets and models. Speedups were particularly dramatic for kNN (e.g., over 100x faster training on GIDS and CICIoV2024) and substantial for RFC (e.g., 35x on OTIDS, 45x on GIDS, 11x on CICIoV2024). While cuML's speed often came with slightly lower accuracy/F1 scores for RFC compared to Scikit-learn (particularly on OTIDS and GIDS), performance metrics for kNN, LR, and XGB remained broadly comparable or nearly identical between the two libraries, even achieving high accuracy with cuML on the CICIoV2024 dataset.

These GPU-accelerated speedups are highly beneficial in domains like the Internet of Vehicles (IoV), where efficient model training and prediction are crucial for real-time applications. However, leveraging cuML presents practical challenges. The library currently lacks specific parameters (e.g., class_weight crucial for imbalanced data), and users may face difficulties with environment setup and updates, especially concerning CUDA compatibility.

Considering these trade-offs, a hybrid approach is recommended: utilize GPU acceleration via cuML for computationally intensive tasks like hyperparameter tuning or feature selection where speed is paramount. Subsequently, the final, optimized model can be trained on a CPU using Scikit-learn. This strategy leverages GPU speed during exploration while ensuring the final model deployment benefits from wider accessibility, lower cost (especially in cloud environments without dedicated GPUs), and potentially more robust parameter options of CPU-based libraries.

VI. Conclusions and Future Work

This study investigated the computational efficiency of GPU-accelerated libraries (cuML) compared to CPU-based implementations (scikit-learn) for machine learning models in IoV intrusion detection. Our comprehensive evaluation across three IoV-specific datasets yielded significant insights into the potential of GPU acceleration for security applications.

The results demonstrated substantial performance improvements, with GPU-accelerated models achieving training time reductions ranging from 6.71x to 159.1x and prediction time improvements of 1.09x to 95.98x compared to CPU implementations. Notably, these significant speed improvements were achieved while maintaining comparable accuracy and F1 scores, indicating no compromise in detection reliability. The RFC and KNN models showed impressive enhancements with GPU acceleration, making them promising candidates for real-time IoV security applications. These findings have important implications for IoV security, as faster training and prediction times enable more frequent model updates and quicker responses to emerging threats. However, challenges remain in implementing these solutions in resource-constrained environments and ensuring scalability across different network sizes.

Future research should focus on developing hybrid CPU-GPU architectures, investigating federated learning approaches for distributed networks, and creating standardized benchmarking frameworks for IoV security solutions. Additionally, field testing in actual IoV environments and research into adversarial attack resistance will be crucial for practical implementation.

VII. Acknowledgments

This work was supported partially by the European Union in the framework of ERASMUS MUNDUS, Project CyberMACS (Project #101082683) (https://cybermacs.eu).

Declaration of Author Contributions

Furkan Çolhak and Hasan Coşkun contributed comprehensively to all stages of this research, including the study design, implementation, data analysis, and initial manuscript writing. Tedi Hoxa and Tsafac Nkombong Regine Cyrille focused on contributing to the writing and refinement of the manuscript. Mert İlhan Ecevit and Mehmet Nafiz Aydın served as supervisors, providing guidance and oversight for the project. Artificial intelligence tools assisted with grammatical corrections and language refinement during manuscript preparation. All authors have read and agreed to the published version of the manuscript.

References

[1] S. Nižetić, P. Šolić, D. L.-d.-I. Gonzalez-De, L. Patrono, *et al.*, "Internet of things (iot): Opportunities, issues and challenges towards a smart and sustainable future," *Journal of cleaner production*, vol. 274, p. 122877, 2020.

[2] W. Xu, H. Zhou, N. Cheng, F. Lyu, W. Shi, J. Chen, and X. Shen, "Internet of vehicles in big data era," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 1, pp. 19–35, 2017.

[3] E. C. Pinto Neto, H. Taslimasa, S. Dadkhah, S. Iqbal, P. Xiong, T. Rahman, and A. A. Ghorbani, "Ciciov2024: Advancing realistic ids approaches against dos and spoofing attack in iov can bus," *Journal of Computer Networks and Communications*, 2024. Accessed: 2024-12-30.

[4] C. Miller and C. Valasek, "Remote exploitation of an unaltered passenger vehicle," in *Black Hat USA*, p. 91, 2015.

[5] P. Beuth, Flüpke, M. Hoppenstedt, M. Kreil, M. Rosenbach, and R. Wilkin, "Volkswagen data breach: Conclusions about the lives of people behind the wheel," 2024. Accessed: 2025-01-05.

[6] S. Purohit and M. Govindarasu, "Ml-based anomaly detection for intra-vehicular can-bus networks," in *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 233–238, 2022.

[7] O. Saber and T. Mazri, "In-vehicle intrusion detection based on machine learning," in *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pp. 1–6, 2023.

[8] J. Liu and W. Fan, "A machine learning-based intrusion detection approach for intelligent connected vehicles," in *2023 24st Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp. 231–234, 2023.

[9] T.-T.-H. Le *et al.*, "Enhancing intrusion detection and explanations for imbalanced vehicle can network data," in *Proceedings of the 12th International Symposium on Information and Communication Technology*, 2023.

[10] E. Limouchi and F. Chan, "Optimized machine learning-based intrusion detection system for internet of vehicles," in *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1151–1157, 2023.

[11] A. B. Kumar and M. Singh, "Enhancing intrusion detection in autonomous vehicles using tree-based machine learning techniques," in *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, pp. 1–7, 2023.

[12] W. Gou, H. Zhang, and R. Zhang, "Multi-classification and tree-based ensemble network for the intrusion detection system in the internet of vehicles," *Sensors*, vol. 23, no. 21, p. 8788, 2023.

[13] S. Kaushik, A. Bhardwaj, and S. S. M. Rahman, "Micord-ids: A hybrid learning system for intrusion detection system for the internet of vehicles," in *2024 7th International Conference on Information and Computer Technologies (ICICT)*, pp. 485–492, 2024.

[14] M. S. Korium, M. Saber, A. Beattie, A. Narayanan, S. Sahoo, and P. H. Nardelli, "Intrusion detection system for cyberattacks in the internet of vehicles environment," *Ad Hoc Networks*, vol. 153, p. 103330, 2024.

[15] S. Bajpai, K. Sharma, and B. K. Chaurasia, "Intrusion detection for internet of vehicles using machine learning," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, 2023.

[16] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Computers and Electrical Engineering*, vol. 105, p. 108543, 2023.

[17] H. Lee, S. H. Jeong, and H. K. Kim, "Otids: A novel intrusion detection system for in-vehicle network by using remote frame," in *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, pp. 57–5709, IEEE, 2017.

[18] E. Seo, H. M. Song, and H. K. Kim, "Gids: Gan based intrusion detection system for in-vehicle network," in *2018 16th annual conference on privacy, security and trust (PST)*, pp. 1–6, IEEE, 2018.

[19] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP)*, vol. 1, pp. 108–116, IEEE, 2018.

[20] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (ddos) attack dataset and taxonomy," in *2019 International Carnahan Conference on Security Technology (ICCST)*, pp. 1–8, 2019.

[21] T. Sutanto, M. Aditya, H. Budiman, M. Ridha, U. Syapotro, and N. Azijah, "Comparison of logistic regression, random forest, svm, knn algorithm for water quality classification based on contaminant parameters," *INTI Journal*, vol. 2022, 11 2024.

[22] J. Price, T. Yamazaki, K. Fujihara, and H. Sone, "Xgboost and support vector machines: Comparing the interpretability of machine learning models," *SSRN Electronic Journal*, 01 2023.

Analysis on features of Human Glaucoma using Human Glaucoma dataset

Sara Trpchevikj
Faculty of Computer Science & Engineering
Saints Cyril and Methodius University
Skopje, North Macedonia
sara.trpchevikj@students.finki.ukim.mk

Petar Sekuloski
Faculty of Computer Science & Engineering
Saints Cyril and Methodius University
Skopje, North Macedonia
petar.sekuloski@finki.ukim.mk

***Abstract*—Eyesight is one of the most important human senses, aiding in our interpretation of the world around us. Therefore, it is necessary that we maintain proper eye health throughout our lives. This paper analyzes relevant data to identify common characteristics of human glaucoma and emphasizes the importance of early detection and treatment of the condition to prevent vision loss.**

***Index Terms*—glaucoma, central corneal thickness, retinal nerve fiber layer, optic neuropathy, intraocular pressure, R**

I. Introduction

Human glaucoma is a group of ocular disorders with multifactorial etiology[1] and with a clinically characteristic intraocular pressure-associated optic neuropathy (ON). ON suggests changes in structural and functional features of the eye, mainly at the optic nerve head (ONH). Over the past 150 years, intraocular pressure (IOP) had been regarded as the primary characteristic of these heterogeneous diseases. However, due to inconsistent measurements among glaucoma patients and healthy individuals, IOP is now regarded as a common risk factor that triggers the disease process [1].

IOP occurs due to excess of aqueous humor fluid, found throughout the inside of the eye, and is controlled by the drainage angle located between the iris and the cornea [2] [3]. However, a compromised drainage system can lead to serious discomfort and impaired vision. Reference [1] notes IOP reduction as the only evidence-based treatment strategy currently known for all types of glaucoma.

The diagnosis is challenging and can lead to failure to identify the condition until significant ON has occurred [4]. Also, clinicians develop a certain degree of belief about the likelihood that a patient has the disease. These differences in data interpretation between experts has given rise to the classification of "glaucoma suspects"— patients for whom insufficient evidence is preventing a definitive diagnosis [1].

In this paper, we present our findings from calculations performed in R using a human glaucoma dataset, with the aim of discussing common characteristics and amplifying the need of routinely eye examinations. Section II discusses related work used for comparison with our own findings. Section III describes the dataset. Section IV describes the calculations and the conclusions gathered from the results. Section V discusses limitations and future developments. Lastly, section VI provides the conclusion of this paper.

[1]The cause, set of causes, or manner of causation of a disease or condition.

II. Related Work

In [4], the feature with the highest usage is the retinal nerve fiber layer 4 (RNFL4) mean, combining 4 separate features directly connected to the RNFL thickness. RNFL is formed by the expansion of the optic nerve fibers and can be described as neural tissue within the eye. For patients with glaucoma, RNFL4 values were lower than the control group, while the values for features like "Ocular pressure" and "Glaucoma Hemifield Test" (GHT) were higher than their healthy counterparts. Difficulty to clinically differentiate normal and glaucomatous eyes was observed due to RNFL thickness being reduced in patients with other ocular disorders.

In [5], the distribution of parameters such as IOP and central corneal thickness (CCT) were determined. The study enrolled 458 patients with glaucoma, including primary open-angle glaucoma (POAG) and normal-tension glaucoma (NTG). Cases with low CCT had lower temporal RNFL thickness, while those with high CCT had smaller optic cup volume. Structural changes and susceptibility to ON show unique variations in patients with different risk factors, most being IOP-independent.

In [6], over the course of a decade, data was collected from 1458 glaucoma patients, with 45.40% of them having primary angle-closure glaucoma (PACG) and 47.19% with secondary glaucoma (SG), respectively. The average age of all patients increased from 56.05 years in 2003 to 57.83 years in 2012. Female gender, cardiovascular disease, and hypertension were associated with PACG. A strong correlation between SG and history of injury and diabetes was also observed.

In [7], CCT correlated positively with the neuro-retinal rim and negatively with visual field loss (VFL). Progressive ON detected in 21% of 861 eyes was statistically independent of CCT. The amount of ON correlated significantly with a thin central cornea, although it may not play a crucial role in the etiology of the disease.

In [8], CCT in glaucoma is thoroughly assessed. The CCT of glaucoma patients and suspects can be influenced my multiple factors, such as ethnicity, age, and the type of glaucoma

treatment. However, correcting IOP did not have any benefit in the assessment of glaucoma, as it was based on structurally normal corneas.

In [9], a peculiarly strong relationship between VFL and RNFL thickness, measured by optic coherent tomography, was detected. Their findings also suggest that the average RNFL thickness is most useful for monitoring glaucoma.

III. Dataset Description

The dataset used for this paper can be found and downloaded from the official site of Kaggle [10]. The dataset was then imported to R as a CSV file. The sample size consists of 499 individuals, i.e., 202 individuals with normal vision (the control group) and 297 individuals with glaucoma. In this paper, we will be analyzing the following features:

A. Glaucoma

This attribute confirms whether or not the individual has been diagnosed with glaucoma based on the examined eye. One of two types of class values may be present for the attribute:

- class 0, meaning the individual has not been diagnosed with glaucoma
- class 1, meaning the individual has been diagnosed with glaucoma.

B. Age

Represents the age of the individual.

C. Ocular pressure

This represents the measured IOP of the patient, where normal values range from 8 to 21 millimeters of mercury ($mmHg$) [11]. Reference [1] accepts the merit of Wilson's "Myth of 21" in reference to an IOP of $21mmHg$ as a value separating the diagnosis of POAG from NTG. However, as the name suggests, the reliability of this cut-off value is and should be questioned. For consistency, we will be referring to the ocular pressure feature as IOP from this point onward in the paper.

D. GHT

The Glaucoma Hemifield Test is an algorithm used by ophthalmologists for automated evaluation of VFL test results in glaucoma [12]. The patient can have one of the following class values for the feature:

- class 0, the patient has not shown significant asymmetric vision
- class 1, the patient achieved vague results
- class 2, the patient has shown significant asymmetric vision.

E. Cornea thickness

It represents the measured thickness of the cornea, in micrometers (μm). We assume that the measurement is most likely done at the highest point of the eye, where the amount of aqueous humor fluid is at its greatest. For consistency, we will be referring to the corneal thickness feature as CCT from this point onward in the paper.

F. RNFL4 Mean

The RNFL4 mean is a synthesized attribute serving as a mean value of 4 combined features related to the RNFL, all measured in micrometers (μm):

- RNFL refers to the thickness of the overall layer
- SUP (superior) refers to the the thickness of the upper part of the RNFL, typically around the top portion of the ONH
- INF (inferior) refers to the thickness of the the lower part of the RNFL, or the bottom portion around the ONH
- TMP (temporal) refers to the thickness of the the side of the RNFL closest to the temple (outer side of the eye, away from the nose) [13].

IV. Calculations and Results

A. Data distribution of age and IOP of patients

As specified in [5], the degree of ON is unique to each patient due to a variety of risk factors, most of which are shown to be independent of their measured IOP. In [4], the mean and standard deviation of the ages of the control group (45.8 ± 16.08 years) are lower than the mean and standard deviation of the ages of glaucoma patients (61.86 ± 13.91 years). This observation led us to wonder about the relationship (or lack thereof) between the age of the patients and their IOP measurements. To assess this connection, we designed a scatter graph of the age and IOP values for both the glaucoma and control group. It is important to separate and compare them for a clearer examination of observable patterns. Identifying such patterns may provide insight into the role of age as a key factor of the disease. The graphs were created using the *ggplot()* function in R, included in the *ggplot2* visualization package.

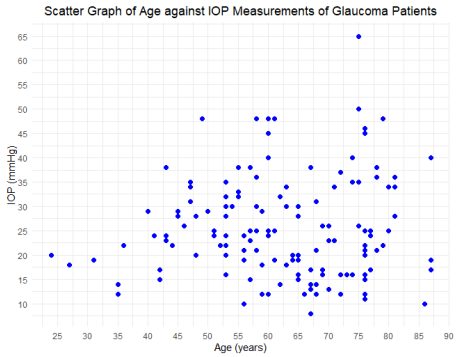


Fig. 1. Scatter graph of age against IOP measurements of glaucoma patients

There are no noticeable patterns or clusters in Fig. 1, possibly indicating a negative relationship between the features for the glaucoma group. This visualization serves as promising evidence of other risk factors having more impact on glaucoma progression. However, the data points seem to create an upward slope on the graph, which validates the popular assumption that older people are significantly more prone to higher IOP.

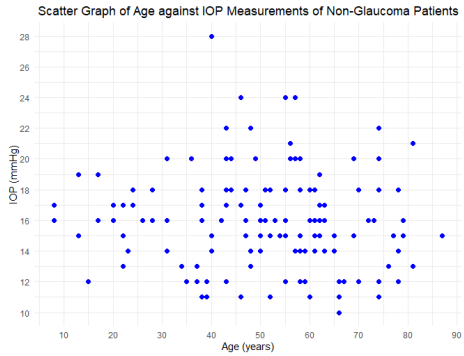


Fig. 2. Scatter graph of age against IOP measurements of non-glaucoma patients

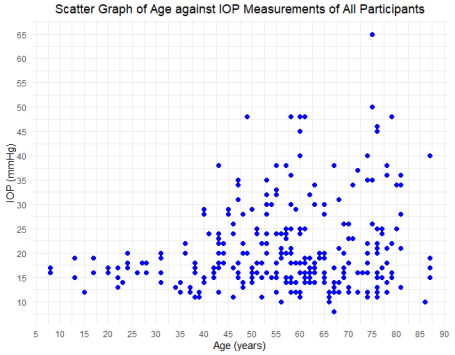


Fig. 3. Scatter graph of age against IOP measurements of all participants

No definite distribution or direction of data points are presents regarding the control group in Fig. 2. Although, lower IOP measurements were recorded in comparison to the glaucoma group, which gives validity, to a certain extent, that people with normal vision are not as susceptible to excessive IOP. Adjustments to the amount of ticks on the axes were made due to this observation.

However, we do notice a trend in the measurements for the entire sample in Fig. 3. A negative skew is prominent, where the left tail is longer and the mass of the distribution is concentrated on the right side of the figure. Majority of the points on the graph are situated at the base of the distribution, following along a straight line. Despite this being a significant observation, it does not suggest a relationship between the age and IOP of the patients. Normal IOP values were recorded for people from 8 years old to 87 years old. Therefore, we cannot confidently affiliate the features.

These observed patterns prompt further speculation regarding the potential relationships between other risk factors and IOP. Continued analysis would involve the construction of similar visualizations based on other demographic or clinical features, enabling comparative evaluation and associations with IOP. Such an approach could offer valuable insights into which factors are crucial for the progression of the diseases.

B. Chi-square test of independence for Glaucoma and GHT

Reference [4] consistently observed high values for features such as IOP and GHT for the glaucoma group. The emergence of a trend in one feature that may suggest or influence the significance of another is relevant for achieving accurate and timely diagnosis of the disease. Understanding these interdependent relationships is essential for enhancing diagnostic precision and improving early intervention strategies. The proportionality of the Glaucoma and GHT features is important to assess because it directly reflects on the predicting power of the GHT algorithm and accurate interpretation of the results. Improving upon the algorithm implies reduced amount of glaucoma suspect diagnoses and increased efforts for developing preventative, rather than mitigative, treatments. We have decided to conduct a Chi-square test of independence for the Glaucoma and GHT features to evaluate their relationship.

The chosen features are qualitative, which describe the status of the diagnosis and result from the GHT. We set the significance level $\alpha = 0.01$ due to the nature of the dataset stemming from a medical laboratory. Stricter significance levels are required to minimize erroneous decisions, where mistakes can be catastrophic. We use the test as a measure of the differences between observed sample frequencies and expected frequencies. We want to test the following hypotheses:

H_0*: "Glaucoma" and "GHT" features are independent*

H_α*: "Glaucoma" and "GHT" features are not independent*

Firstly, a contingency table for the features is constructed using the *table()* function in R, placing the feature columns as our arguments within the brackets. The table allows us to compare the data and observe how it overlaps. Labels are appropriately added using the *rownames()* and *colnames()* functions, and highlighting marginal sums using the *addmargins()* function.

TABLE I
CONTINGENCY TABLE FOR "GLAUCOMA" AND "GHT" FEATURES

	No vision asymmetry	**Unclear**	**With vision asymmetry**	**Total**
No glaucoma	141	26	35	202
Has glaucoma	28	11	258	297
Total	169	37	293	499

To perform the Chi-square test of independence, we use the *chisq.test()* function, passing the contingency table as the argument. The function uses the following formula, with $(r-1)\times(k-1)$ degrees of freedom, r representing the two classes of the "Glaucoma" feature, and k representing the three classes of the "GHT" feature:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} \tag{1}$$

The null hypothesis is rejected if the calculated value of the test statistic is greater than the critical value C, which we can read from the Chi-square distribution table.

$$C = \left(\chi^2_{\alpha,(r-1)\times(k-1)}, +\infty\right) \tag{2}$$

The results from the Chi-square test of independence are the following:

$$\chi^2 = 242.05 \tag{3}$$

$$degrees\ of\ freedom = 2 \tag{4}$$

$$p - value < 2.2 \times 10^{-16} \tag{5}$$

The test statistic value in (3) quantifies the extent of deviation between the observed frequencies in the contingency table and the expected frequencies under the assumption of independence between the two qualitative features. A large value like ours for the test statistic suggests a significant discrepancy between the observed and expected frequencies, indicating a potential association between the two features.

Additionally, the p-value is provided, which is computed using the Chi-square test statistic and the degrees of freedom. The p-value represents the probability that a Chi-square distribution with the specified degrees of freedom would produce a value more extreme than the observed statistic. The value is derived by integrating the probability density function of the Chi-square distribution:

$$p - value = P(\chi^2 \geq observed\ \chi^2) \tag{6}$$

The p-value in (5), an exceedingly small value, is essentially approaching zero. This indicates that the likelihood of observing a χ^2 test statistic equal to or greater than (3), under the assumption of independence between the two features, is virtually negligible. Given such a small p-value, which is well below conventional significance thresholds (e.g., 0.05, 0.01, or even 0.001), we have strong evidence to reject the null hypothesis.

The final step is assessment of the critical domain *C*:

$$\chi^2_{\alpha,(r-1)(k-1)} = \chi^2_{0.01,2} = 9.21 \tag{7}$$

$$C = (9.21, +\infty) \tag{8}$$

Given our p-value in (5), which is smaller than the chosen significance level of 0.01, and considering that the test statistic in (3) falls within the critical region C in (8), we conclude that the null hypothesis is rejected. This implies that the diagnosis of glaucoma is dependent on the outcome of the GHT.

While this analysis demonstrates a significant connection, it does not imply a causal relationship between the features and other underlying factors could influence the results. Also, limitations such as the absence of data and the applicability of the findings to other populations should be considered. To further explore the association, future research could include a more detailed analysis incorporating additional risk factors, such as gender and lifestyle habits which may provide deeper insights into how these factors influence the development of glaucoma. Expanding the scope of the study to larger and more diverse datasets would further strengthen the evidence and enhance clinical relevance of the findings.

C. Regression and correlation analysis of RNFL4 mean and CCT

Reference [4] praised the newly synthesized RNFL4 feature due to its proximity to the optic nerve. In [5], the findings suggest great damage being done to the RNFL and/or limit the amount of light information being transmitted to the brain, with CCT being a key factor of these disruptions. Reference [7] identified CCT values correlating negatively with loss of the patient's visual span. On the other hand, independence of CCT and ON progression was confirmed for 21% of eyes included in the study. In [8], the causes for low CCT measurements were attributed to a variety of demographic and clinical factors, emphasizing the obscure nature of the diseases. In [9], the researchers credited the RNFL as the most useful aspect of monitoring glaucomatous eyes.

The dynamic between the CCT and RNFL measurements is important to analyze because of their use by clinicians, without the need for custom algorithms. We chose to inspect the connection using a regression and correlation analysis.

Regression analysis is used to determine the type of functional relationship between features and to predict or estimate the value of one variable for a given value of the other variable. It serves to view how a dependent variable (RNFL4 mean) changes when one or more independent variables (CCT) change. Correlation analysis, on the other hand, is related to measuring the strength of the relationship between variables, i.e., the degree of correlation.

Changes in CCT are associated with or correlate to changes in RNFL. CCT is often considered an anatomical measurement that can be predictive of changes or damage in the RNFL. Also, RNFL thickness is frequently treated as a response when studying the progression of damage to the optic nerve.

The initial step is to construct a mathematical model that expresses the RNFL4 mean value as a function of the CCT measurements, using the database of RNFL4 mean measurements at specific CCT values. In order to determine the type of model, we constructed a simple scatter graph for the features to observe the placement and direction of the data points. We placed the independent variable on the x-axis, and the dependent variable on the y-axis. The graph was constructed using the *ggplot()* function in R, included in the *ggplot2* visualization package.

Fig. 4 shows no clear linear trend between RNFL4 mean and CCT values. The points are widely scattered, and the cloud of points looks roughly horizontal across the range of CCT values. There is no obvious upward or downward slope, and no clear curvilinear (e.g., quadratic or exponential) pattern either. There may be no statistically or clinically significant association between the features. However, we fitted a linear regression first to formally assess the relationship.

The statistical model we use to relate the dependent variable *Y* to the independent variable *X* is a linear regression model.

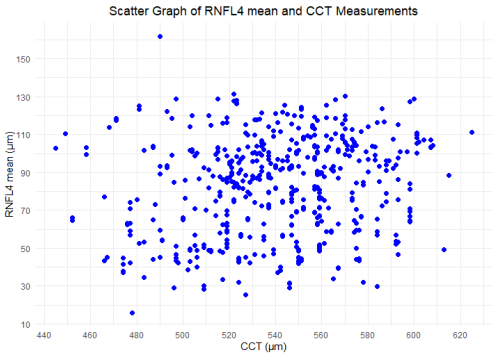


Fig. 4. Scatter graph of RNFL4 mean and CCT measurements

This model specifies the mathematical expectation of the random variable *Y*, given the independent variable $X = x$, as a linear function of a set of unknown parameters. The linear regression model is the following:

$$E(Y|x) = \beta_0 + \beta_1 \times x \tag{9}$$

The regression model in (9) connects *Y* with the value of a independent variable *X* and proposes that the expected value of *Y*, for a given value *X*=*x*, is a line when displayed graphically.

Next, we introduced some numerical measures that will support our observations from Fig. 4. Such a measure is the correlation coefficient (r). The correlation coefficient measures the direction and strength of the linear relationship between our two quantitative features. It is done using the *cor()* function in R and uses the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}} \tag{10}$$

$$r = 0.1745266 \tag{11}$$

The correlation coefficient result in (11) indicates a weak linear relationship between the two features. While the value of *r* is close to zero, suggesting that the linear association is not strong, the positive sign of the coefficient implies that as one feature increases, there is a slight tendency for the other feature to increase as well. This association provides partial support for the findings in [5], as it suggests that the relationship they observed between these variables is present in our data, although not robust or predictive. It is important to note that such a low correlation coefficient limits the practical and clinical significance of this relationship, suggesting other factors may play a more substantial role in the variability.

Eq. (9) represents a simple linear regression model. β_0 represents the y-intercept, while β_1 is the slope coefficient, and can be interpreted as the change in the probability that $Y = 1$ resulting from a unit increase in x when the remaining independent variables (if any) are held constant [14].

To find the linear regression equation in R, we used the *lm()* function, where RNFL4 mean values are dependent on the CCT values ($Y \sim X$). The resulting equation is the following:

$$Y = 10.1894 + 0.1339 \times x \tag{12}$$

The linear regression equation in (12) characterizes the relationship between the dependent variable Y and the predictor variable x. Specifically, the intercept $\beta_0 = 10.1894$ represents the estimated value of Y when $x = 0$, as starting point for the regression line on the y-axis. The slope coefficient $\beta_1 = 0.1339$ indicates that for every one-unit increase in x, the predicted value of Y increases by approximately $0.1339\mu m$. This positive slope reflects a subtle association between the two variables.

Using the equation in (12), we can place the regression line over the scatter graph in Fig. 4 to assess if the features have a linear or non-linear relationship. This was done by expansion of the arguments with the result in (12).

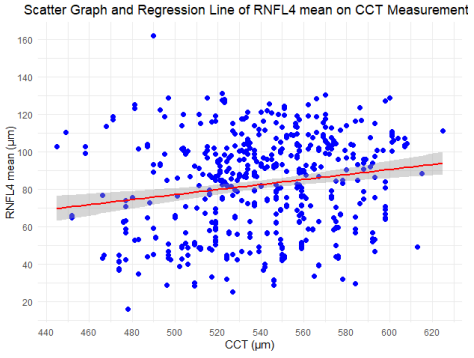


Fig. 5. Scatter graph and regression line of RNFL4 mean on CCT measurements

In Fig. 5, the data points show substantial variability across the range of CCT values, with RNFL4 mean measurements spanning $16 - 162\mu m$. The regression line indicates a slight positive trend, suggesting that as CCT increases, RNFL4 mean also tends to increase, although the effect size still appears modest. The confidence band around the regression line remains relatively narrow, reflecting some stability in the linear estimate. However, there is a wide spread of data points around the regression line, suggesting that CCT explains only a small proportion of the variance in the RNFL4 mean. Overall, the considerable scatter in the data suggests that CCT alone is likely an insufficient predictor of RNFL4 mean without considering additional factors.

To further assess the model's fit, the coefficient of determination (R^2) is deduced for the proportion of variance in Y explained by X. It represents the measure of the spread of observations about the regression line. For linear models, R^2 is between 0 and 1. A larger R^2 value indicates that the data is tightly packed around the regression line, while a lower value suggest high variability and noisiness of the data.

Coefficient of determination is defined as the fraction of the total sum of squares (SST) and the residual sum of squares (SSR), where $\bar{y}$ represents the mean of the dependent variable Y, obtained using the *mean()* function in R, and $\hat{y}$ are the

predicted values of Y in regards to the linear model, obtained using the *predict()* function in R.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 330971.9 \tag{13}$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 10081.25 \tag{14}$$

$$R^2 = \frac{SSR}{SST} = \frac{10081.25}{330971.9} = 0.0305 \tag{15}$$

The coefficient of determination in (15) is close to zero. This indicates that only about 3.05% of the total variability in the dependent variable Y can be explained by the linear regression model using the predictor X. Despite the small R^2 value, it is important to interpret this result carefully. A low R^2 does not automatically mean that the model is a good fit. Rather, it typically suggests that the model explains very little for the variation in the data. In this case, the low R^2 points to a high degree of scatter around the regression line, seen in Fig. 5, and indicates that there is little to no linear dependence between the two features. Additional diagnostics, such as residual analysis and consideration of alternative models, may be needed to determine whether a linear approach is suitable or a more complex, nonlinear relationship should be explored.

V. Limitations and Developments

A flaw in the organization of the dataset is the lack of classification of the type of glaucoma each patient has. Rather, they are all grouped as having a certain, yet unknown type of glaucoma or not having the disease entirely. To improve treatment strategies and limit the amount of glaucoma suspects, detailed documentation should be provided for the kind of glaucoma being diagnosed. Also, including VFL as a complementary feature would have been of value to the research for more elaborate and detailed conclusions, specifically regarding GHT from which the diagnosis is confirmed.

The lack of demographic features is prominent in this particular dataset, and expanding it with a variety of information such as gender, geographical location, lifestyle habits, and even residential pollution levels could impact the results.

Assessment of the potential relationships between IOP and other risk factors is mandatory for future developments. Visualizations on demographic and clinical features could spark discussion about the delicate characteristic that is IOP.

Underlying factors influencing both a glaucoma diagnosis and GHT results are highly plausible and should be evaluated. Limitations and lack of applicability of the findings can exist to other populations.

Despite the linear regression model suiting the RNFL4 and CCT features to a certain degree, conducting a residual analysis and testing curvilinear models alongside could be of use. With data distributions as complicated as these to correctly interpret, experimenting to assess the fit of other models is key for future developments.

VI. Conclusion

The analysis of the scatter graphs of age and IOP measurements across the glaucoma, control, and combined groups shows no clear or consistent relationship between age and IOP. The results of the Chi-square test of independence provide strong statistical evidence that the diagnosis of glaucoma is significantly associated with the outcome of the GHT. The observed distribution of glaucoma diagnoses across GHT categories differs from what would be expected under the assumption of independence. Our regression and correlation analysis between RNFL4 mean and CCT measurements revealed a weak, but positive association. While the linear regression model identified a small positive trend, the wide scatter of data points suggests that CCT alone is not a strong predictor of RNFL4 mean in our dataset. Understanding the influences of the diseases can significantly impact ophthalmology, and this analysis has given us insight into glaucoma, all with the aim of improving knowledge and developing preventative medicine.

References

[1] Casson, R.J., Chidlow, G., Wood, J.P., Crowston, J.G. and Goldberg, I., "Definition of glaucoma: clinical and experimental concepts. Clinical & Experimental Ophthalmology", pp. 341-349, February 2012. [Online]. Available: https://doi.org/10.1111/j.1442-9071.2012.02773.x.

[2] [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK553209/. [Accessed: Feb. 16, 2025]

[3] [Online]. Available: https://www.aao.org/eye-health/diseases/what-is-glaucoma. [Accessed: Feb. 16, 2025]

[4] Kim, S.J.; Cho, K.J.; Oh, S, "Development of machine learning models for diagnosis of glaucoma", May 2017. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177726.

[5] Omodaka, K., Kikawa, T., Kabakura, S. et al., "Clinical characteristics of glaucoma patients with various risk factors", BMC Ophthalmol 22, 373, September 2022. [Online]. Available: https://doi.org/10.1186/s12886-022-02587-5

[6] Zhao, Y; Fu, J.L.; Li, Y.L.; Li, P; Lou, F.L., "Epidemiology and clinical characteristics of patients with glaucoma: An analysis of hospital data between 2003 and 2012", Indian Journal of Ophthalmology, pp. 825-831, November 2015. [Online]. Available: https://journals.lww.com/ijo/fulltext/2015/63110/epidemiology_and_clinical_characteristics_of.4.aspx.

[7] Jonas, J. B.; Stroux, A.; Velten, I.; Juenemann, A.; Martus, P.; Budde, W. M., "Central Corneal Thickness Correlated with Glaucoma Damage and Rate of Progression", Invest. Ophthalmol. Vis. Sci.; pp. 1269-1274, April 2005. [Online]. Available: https://doi.org/10.1167/iovs.04-0265.

[8] Belovay, G.W.; Goldberg, I., "The thick and thin of the central corneal thickness in glaucoma", February 2018. [Online]. Available: https://doi.org/10.1038/s41433-018-0033-3.

[9] Kanamori, A.; Nakamura, M.;Escano, M.F.T; Seya, R.; Maeda, H.; Negi, A., "Evaluation of the glaucomatous damage on retinal nerve fiber layer thickness measured by optical coherence tomography", American Journal of Ophthalmology, Volume 135, Issue 4, pp. 513-520, March 2003. [Online]. Available: https://doi.org/10.1016/S0002-9394(02)02003-2.

[10] [Online]. Available: https://www.kaggle.com/datasets/gchan357/human-glaucoma. [Accessed: Jan. 24, 2025]

[11] [Online]. Available: https://sistinaoftalmologija.mk/glaukom/. [Accessed: Feb. 18, 2025]

[12] Åsman, P; Heijl, A.; "Glaucoma Hemifield Test: Automated Visual Field Evaluation" Arch Ophthalmol., 110(6), pp. 812–819, June 1992. [Online]. Available: https://jamanetwork.com/journals/jamaophthalmology/article-abstract/639714.

[13] Çömez, A.T.; Eser, İ.; Bakar, C., "Is Single Measurement Enough to Get a Reliable Result with Optical Coherence Tomography?", Turkish Journal of Ophthalmology, 42(1), pp. 11-15, December 2011. [Online]. Available: https://oftalmoloji.org/articles/doi/tjo.42.680770.

[14] [Online]. Available: https://www.sciencedirect.com/topics/mathematics/slope-coefficient [Accessed: Feb. 26, 2025]

Interactive Recommendation System with Graph Neural Networks

David Arsovski
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
Skopje, North Macedonia
david.arsovski@students.finki.ukim.mk

Dimitar Trajanov
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
Skopje, North Macedonia
dimitar.trajanov@finki.ukim.mk

Marija Saveska
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
Skopje, North Macedonia
marija.saveska@students.finki.ukim.mk

***Abstract*—Modern software development advances have increased user expectations for personalized experiences, highlighting the demand for recommendation systems that are precise, flexible, and context-aware. To address this, we've developed an interactive book recommendation system utilizing extensive user-generated data and powered by Graph Neural Networks (GNNs). The system incorporates architectures like GraphSAGE, Graph Attention Networks (GAT), and a Transformer-based model, which are enhanced with RoBERTa embeddings to capture complex user-product relationships. Our evaluations indicate the GAT model outperforms GraphSAGE and the Transformer-based model. We also built a web tool that enables an interactive recommendation feedback loop, demonstrating the practical use of such systems.**

***Index Terms*—Graph Neural Networks (GNN), Transformer Architecture, Recommendation Systems**

I. Introduction

The rapid expansion of content in the digital era makes finding relevant items a challenge, highlighting the need for robust recommendation systems. These systems solve this by suggesting items based on user interests and previous interactions, which enhances the user experience and can inform business strategies.

Although traditional methods such as collaborative and content-based filtering [1] have been effective, they often encounter difficulties with scalability, data sparsity, and the cold start problem. Recent advances in deep learning, more specifically Graph Neural Networks, offer promising solutions by modeling users and products as nodes and their interactions as edges. This modeling captures both direct and higher-order relationships, leading to more effective learning of user preferences.

This paper introduces an interactive recommendation system built with Graph Neural Networks (GNNs) that adjusts its suggestions based on real-time user feedback. We cover the system design, its implementation and how it was evaluated, discuss technical hurdles and show the advantages of using GNNs for recommendation tasks. Key contributions are a comparative analysis of three GNN architectures and a new web-based tool designed for interactive user engagement.

II. Related Work

The evolution of recommendation systems can be traced back to the pioneering work of Elaine Rich in 1979, titled "User Modeling via Stereotypes" [2]. Rich introduced the concept of classifying users by employing generalized models or 'stereotypes' to quickly categorize new users with limited initial data. As users interact more with the system, the models are refined to create more personalized profiles. This approach, which balances general assumptions and tailored user profiles, makes the groundwork for future developments in personalized recommendation systems.

The influence of graph-based methods is also clear in how these systems have developed. One of the earliest graph-based models is the PageRank algorithm, developed by Larry Page and Sergey Brin in 1998 [3]. Initially designed to rank web pages for search engines, PageRank operates on the principle that a node's significance is determined by its connections to other important nodes. Although its primary application was web search, the underlying idea has influenced various domains, including academic research and author classification, due to its effective way of quantifying relative importance within a graph structure [4].

Traditional recommendation techniques primarily include collaborative filtering [5] and content-based filtering. Collaborative filtering is further subdivided into user-based and item-based approaches. User-based collaborative filtering identifies users with similar preferences, whereas item-based filtering analyses the relationships between items to suggest similar products [6]. While these methods have proven effective, they face challenges such as scalability with large user-product matrices, data sparsity, and the cold-start problem, where insufficient data hampers accurate recommendations.

Content-based filtering [7], on the other hand, relies on the attributes of products—such as titles, authors, descriptions and genres to create recommendation models. By converting these attributes into vectorized forms, techniques like TF-IDF have been employed to gauge the significance of different terms. Although this method is computationally efficient and particularly effective for domains like book recommendations [8], it often falls short in capturing the more complex, latent

relationships present in larger datasets.

To overcome the limitations of individual approaches, hybrid systems emerged. These systems combine collaborative filtering, content-based methods, and other techniques to leverage the strengths of each while mitigating their weaknesses [9]. A notable example of this integration is evidenced by the Netflix Prize competition [10]. Initiated in 2006, the competition challenged participants to improve movie recommendation accuracy by at least 10% using ensemble models. The winning team, BellKor [11], combined multiple methods, including matrix factorization techniques like Singular Value Decomposition (SVD), to address local and global relationships within the data, thereby enhancing the model's performance.

In recent years, Graph Neural Networks have gained prominence as a promising solution for recommendation systems [12]. GNNs are particularly adept at modeling complex relationships in data by representing users and products as nodes and their interactions as edges in a graph structure. This allows the system to capture both direct connections and higher-order, indirect interactions (neighbors of neighbors), which can significantly improve the accuracy of recommendations. Several surveys provide overviews of GNN applications in recommender systems [13]. The ability of GNNs to integrate information from diverse sources and model complex interdependencies has positioned them as a cutting-edge approach in the field of artificial intelligence and recommendation systems.

The trajectory from early user modeling to sophisticated hybrid systems and now to advanced GNN-based methods highlights the ongoing efforts to improve personalization and accuracy in recommendation systems. The integration of graph structures and neural network techniques promises to address many of the scalability and sparsity challenges faced by traditional methods, paving the way for more robust and context-aware recommendations.

III. Graph Neural Networks for Link Prediction in Recommendation Systems

The evolution of recommendation systems increasingly relies on deep learning techniques, particularly Graph Neural Networks - to process the relational data inherent in user–product interactions. In this work, we construct and compare three GNN architectures for link prediction, aiming to determine whether a user would rate a book with a score of 3 or higher (out of 5). The three models are based on GraphSAGE, Transformer-based GNNs, and Graph Attention Networks (GAT).

A. Data Collection and Preprocessing

We utilize publicly available datasets from Amazon and the Google Books API, initially containing approximately 200,000 books and 30 million reviews. To ensure data quality and manage computational constraints, we:

- Remove entries with missing or duplicate titles.
- Select users with at least 10 reviews and books with at least 5 reviews.
- Exclude books with multiple authors or ambiguous genres.

After filtering, the dataset comprises roughly 12,000 book nodes, 25,000 user nodes, and 500,000 edges. Figure 1 shows a subset of the graph where popular books and prolific reviewers are easily identifiable. Due to the high density of interactions in the full graph, which is difficult to render clearly, this subset aims to illustrate the nature of user-book connections where popular items and active users form denser clusters.

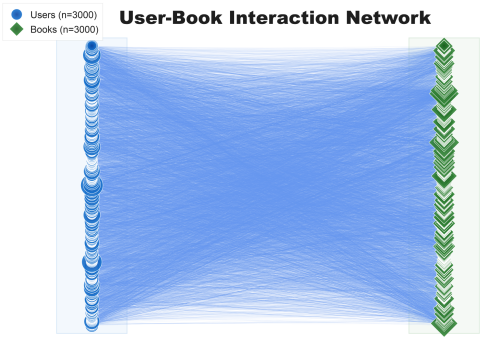


Fig. 1: Visualizing the links between a subset (n=3000) of user and book nodes.

B. Feature Engineering

For categorical attributes such as publisher, author, and genre, we apply a Label Encoding scheme to convert textual information into numerical labels. To capture the semantic richness of book titles and descriptions, we use the RoBERTa model to generate vector embeddings [14]. However, due to the high dimensionality of these embeddings, we apply Principal Component Analysis (PCA) to reduce the vectors to a manageable size (4 dimensions for titles and 16 dimensions for descriptions), thus retaining critical semantic features while enhancing computational efficiency.

C. Model Architectures

A recommendation system using a Graph Neural Network must effectively capture relationships between books and users by leveraging graph structure and node attributes. Since our data is in graph form, nodes are converted into vector representations, known as node embeddings. We use PyTorch and PyTorch Geometric for building the GNNs.

a) GraphSAGE Model [15]: GraphSAGE utilizes *Graph Sample and Aggregate* convolutional layers to aggregate neighbor features. It is scalable as it samples a fixed number of neighbors, enabling efficient operation on large graphs. Our model consists of two `SAGEConv` layers with a `Dropout` layer in between to prevent overfitting. Given the high density of our graph, excessive layers could lead to over-smoothing, making distant nodes appear too similar. We apply the `ReLU` activation function for regularization.

```
class GraphSAGE_3(torch.nn.Module):
```

```
    def __init__(self, hidden_channels, dropout=0.3):
        super().__init__()
        self.conv1 = SAGEConv(hidden_channels, hidden_channels)
        self.dropout1 = nn.Dropout(dropout)
        self.conv2 = SAGEConv(hidden_channels, hidden_channels)

    def forward(self, x, edge_index):
        x = self.conv1(x, edge_index).relu()
        x = self.dropout1(x)
        return self.conv2(x, edge_index)
```

Listing 1: GraphSAGE model implementation.

To predict interactions, we compute node embeddings for users and books and apply a dot product to classify interactions:

```
class Classifier(torch.nn.Module):
    def forward(self, x_user: Tensor, x_book: Tensor, edge_label_index: Tensor) -> Tensor:
        edge_feat_user = x_user[edge_label_index[0]].to(dtype=x_book.dtype)
        edge_feat_book = x_book[edge_label_index[1]]
        return (edge_feat_user * edge_feat_book).sum(dim=-1)
```

Listing 2: Classifier for user-book interactions.

We dynamically update embeddings during training to enrich data representations.

b) Transformer-based Model [16]: Transformers have revolutionized AI, particularly in NLP, with architectures like GPT [17]. Using *self-attention*, they capture complex dependencies and long-range relationships in data.

Our model includes two `TransformerConv` layers that dynamically weigh neighboring nodes based on learned attention scores. This approach captures both local and global graph structures but comes with increased computational complexity, making training and inference slower on large graphs.

c) Graph Attention Network (GAT) Model [18]: GAT improves upon traditional GNNs by incorporating *attention mechanisms*. Unlike GraphSAGE, which treats all neighbors equally, GAT assigns *attention scores* to weigh each neighbor's importance. This allows the model to prioritize crucial connections, enhancing representation learning.

Like our previous models, we use two `GATConv` layers, where attention coefficients determine the influence of each neighboring node. While GAT excels at modeling local relationships, Transformers provide a broader understanding of both local and global dependencies, albeit at a higher computational cost.

D. Model Training

We optimize model parameters to improve prediction accuracy and relevance. For our graph neural network, the graph is segmented into training, validation, and test sets using a method that splits the edges rather than the nodes. This ensures that edges in the training set do not overlap with those in the validation or test sets, and similarly, validation edges are distinct from test edges.

Specifically, a function such as `RandomLinkSplit` is used to randomly allocate 10% of the edges for validation and another 10% for testing, with the remaining edges used for training. Additionally, a subset of training edges is reserved for supervision (with the rest used for message propagation), and negative edges—non-existent connections—are generated at a specified ratio (2:1) to help the model learn the difference between real links and random ones.

Due to the large size of the graph, we employ a data loader (e.g., `LinkNeighborLoader`) that generates mini-batches by sampling a fixed number of neighbors per node. This method reduces memory usage by processing smaller subgraphs iteratively, though it may omit some contextual information from unselected neighbors.

We use the Adam optimizer [19] and Binary Cross-Entropy loss for binary edge prediction. In each epoch, the model processes these mini-batches, applies a sigmoid to its outputs to produce probabilities in the range $[0, 1]$, computes the loss, and updates the weights via backpropagation. Given the model's complexity and potential for overfitting, training is conducted over a small number of epochs to minimize the average loss.

IV. Evaluation

Evaluating our recommendation model is the final step, and we will assess performance using several standard metrics suited for link prediction and recommendation tasks.

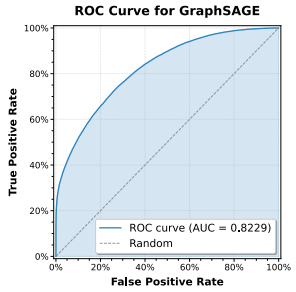
1) ROC-AUC: This metric measures the model's ability to distinguish between positive and negative links. The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR). A higher Area Under the Curve (AUC) indicates a better model at distinguishing between positive and negative classes. Figure 2 shows the ROC curves for the GraphSAGE, Transformer, and GAT models.

2) RMSE: The Root Mean Square Error (RMSE) quantifies the difference between predicted and actual values, placing extra emphasis on larger errors:

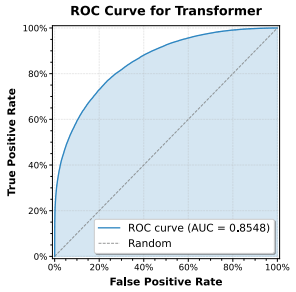
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where n is the number of edges, y_i is the actual state (1 if the edge exists, 0 otherwise), and $\hat{y}_i$ is the predicted probability of interaction. Lower RMSE values indicate higher model accuracy and stability. While RMSE is traditionally used for regression tasks, we adapt it here to measure the deviation between predicted interaction probabilities (continuous) and binary edge labels (0/1). This provides additional insight, penalizing large errors in confidence scores (e.g., predicting 0.9 for a non-existent edge). Figure 3 shows the RMSE comparison.

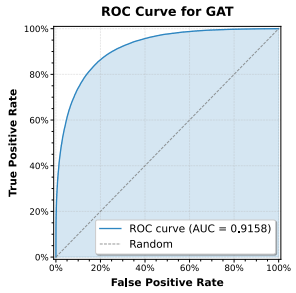
3) Hit@k: Hit@k measures the fraction of the top k recommendations that match the actual user preferences. It directly reflects user experience, as recommendations are typically judged by top suggestions. After sorting predictions by



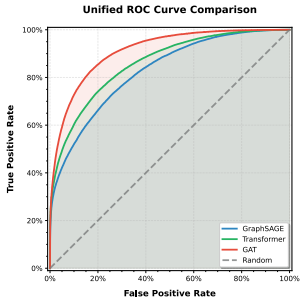
(a) ROC Curve for the GraphSAGE model.



(b) ROC Curve for the Transformer model.



(c) ROC Curve for the Graph Attention Network model.



(d) ROC Curve comparison between models.

Fig. 2: ROC Curves for different graph-based models.

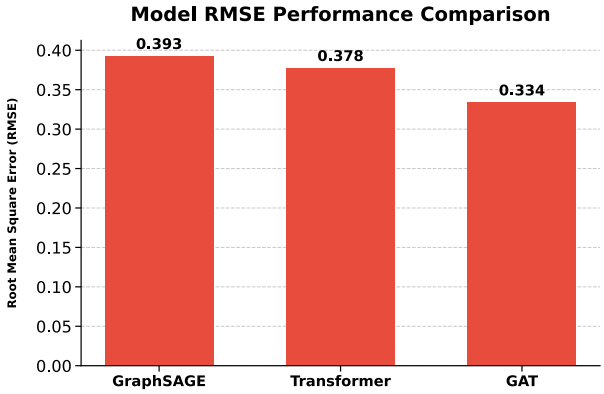


Fig. 3: RMSE Performance Comparison.

probability, we count how many of the top k correspond to existing edges:

$$hit@k = \frac{\text{Number of hits in top } k}{k}$$

This metric is especially relevant in recommendation systems, where users focus on the highest-ranked suggestions. Figure 4 compares the Hit@k performance for different values of k.

A. Overall Performance Analysis

The performance of the three GNN models—GraphSAGE, Transformer, and GAT—was evaluated using the ROC-AUC, RMSE, and Hit@k metrics, as detailed above. Table I summarizes these key evaluation metrics.

The GAT model consistently achieved the best scores across all metrics, demonstrating its superior capability in this book recommendation task. Specifically, GAT yielded the highest ROC-AUC (0.91) and Hit@50k (0.70), and the lowest RMSE (0.334). While GraphSAGE offers faster training and Transformers can capture long-range dependencies, such dependencies were less prominent in our dataset, making GAT's attention mechanism more effective for capturing influential local neighborhood information. Based on these evaluations, we adopt the GAT model for book recommendations in our interactive tool.

Model	ROC-AUC	RMSE	Hit@50k
GraphSAGE	0.82	0.393	0.60
Transformer	0.85	0.378	0.68
GAT	**0.91**	**0.334**	**0.70**

TABLE I: Evaluation Metrics Comparison.

V. Interactive Web Page

To demonstrate the practical applicability and dynamic nature of the GNN-based recommendations, particularly the best-performing GAT model, we developed an interactive web page. It enables real-time exploration of how a real user would engage with our model and how the model adapts to user preferences. It features an interface for browsing and rating books and a backend server that processes requests and delivers personalized recommendations dynamically. The goal is to create an intuitive experience where users seamlessly explore model-generated suggestions, increasing engagement by presenting relevant and appealing book recommendations.

A. System Architecture and Implementation

The frontend uses **Svelte/SvelteKit** for its ease of use, and **DaisyUI** for rapid UI development. The backend employs **Flask** due to its Python compatibility, simplicity, and seamless integration with our PyTorch-based recommendation model. This combination ensures fast loading, efficient data transfer, and responsive user experiences. The architecture of the system can be seen in fig. 5.

1) Key Technologies:

- **SvelteKit** enables server-side rendering (SSR) for quick initial page loads and dynamic routing
- **Flask API** handles three primary endpoints:
  - `/api/recommend`: Returns 10 personalized book recommendations
  - `/api/book`: Provides metadata for a specific book ID
  - `/api/search`: Processes text queries (titles/authors) and returns matching books

Our pre-trained **Graph Attention Network** model (selected for superior performance in earlier evaluations) is loaded into Flask's memory during initialization. It undergoes fine-tuning using real-time user interaction data.

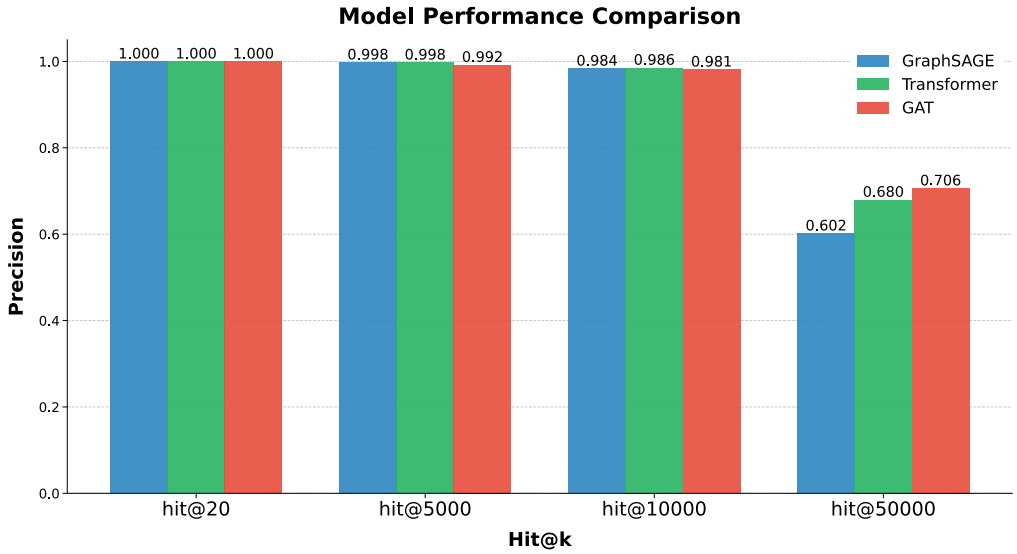


Fig. 4: Hit@k Model Performance Comparison.

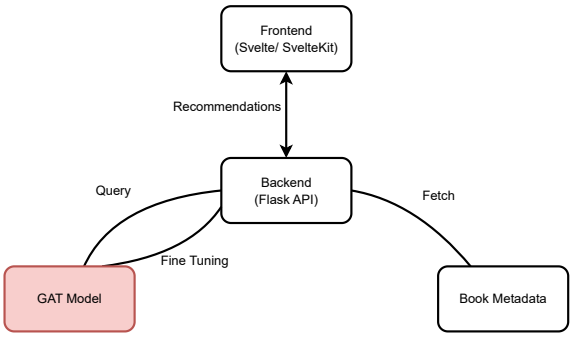


Fig. 5: System Architecture Diagram.

B. Recommendation Process

- **Onboarding:** Users select preferred genres to initialize recommendations. Preferences are stored locally to avoid redundant prompts.
- **Rating Interface:** An intuitive card-based interface allows users to rate books (1–5 stars). Ratings are stored locally and submitted to the backend.
- **Search Functionality:** A search bar facilitates exploration of ∼10,000 books by title/author.
- **Dynamic Recommendation Generation:** User ratings are used to fine-tune the model in real time. New user nodes are added to the graph, and recommendations are filtered for diversity using cosine similarity.

VI. Conclusion

In this paper, we investigated graph-based recommendation models for books by leveraging modern graph neural network architectures. We developed and compared three distinct models—GraphSAGE, Transformers, and Graph Attention Networks—to assess their performance in capturing complex relationships in the dataset.

Our evaluation was based on metrics that measured the accuracy of predictions and the relevance of the top recommendations. Despite the inherent challenges posed by limited and lower-quality data—common in academic or small-scale studies—the results indicate that even with a relatively small dataset, these graph neural network models can achieve high precision and deliver adequate recommendations.

Although Graph Attention Networks (GAT) demonstrate strong performance, effectively addressing the cold-start problem for new users and items remains a significant challenge. To mitigate this, our future research will focus on integrating more content-based embeddings, such as those derived from RoBERTa, to develop a hybrid recommendation system. All data was utilized in strict compliance with Amazon and Google's research terms, and user privacy was maintained through the use of anonymized identifiers.

Our findings confirm the potential of graph neural network-based models to transform recommendation systems, even in scenarios with constrained data resources, and set the stage for further research aimed at optimizing their performance in real-world applications. The source code is publicly available to encourage further research [20].

References

[1] S. B. Poonam B. Thorat, R. M. Goudar, "Survey on collaborative filtering, content-based filtering and hybrid recommendation system," *International Journal of Computer Applications*, vol. 110, no. 4, pp. 31–36, January 2015. [Online]. Available: https://ijcaonline.org/archives/volume110/number4/19308-0760/

[2] E. Rich, "User modeling via stereotypes," *Cognitive Science*, vol. 3, no. 4, pp. 329–354, 1979.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[4] D. F. Gleich, "Pagerank beyond the web," *SIAM Rev.*, vol. 57, no. 3, p. 321–363, Jan. 2015. [Online]. Available: https://doi.org/10.1137/140976649

[5] Y. Koren and R. Bell, *Advances in Collaborative Filtering*. Boston, MA: Springer US, 2011, pp. 145–186. [Online]. Available: https://doi.org/10.1007/978-0-387-85820-3_5

[6] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th International Conference on World Wide Web*, ser. WWW '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 285–295. [Online]. Available: https://doi.org/10.1145/371920.372071

[7] R. Van Meteren and M. Van Someren, "Using content-based filtering for recommendation," in *Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop*, vol. 30. Barcelona, 2000, pp. 47–56.

[8] P. Mathew, B. Kuriakose, and V. Hegde, "Book recommendation system through content based and collaborative filtering method," in *2016 International conference on data mining and advanced computing (SAPIENCE)*. IEEE, 2016, pp. 47–52.

[9] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[10] J. Bennett and S. Lanning, "The netflix prize," 2007.

[11] R. M. Bell and Y. Koren, "Lessons from the netflix prize challenge," *Acm Sigkdd Explorations Newsletter*, vol. 9, no. 2, pp. 75–79, 2007.

[12] Q. Zhang, J. Lu, and Y. Jin, "Artificial intelligence in recommender systems," *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 439–457, 2021.

[13] C. Gao, Y. Zheng, N. Li, Y. Li, Y. Qin, J. Piao, Y. Quan, J. Chang, D. Jin, X. He *et al.*, "A survey of graph neural networks for recommender systems: Challenges, methods, and directions," *ACM Transactions on Recommender Systems*, vol. 1, no. 1, pp. 1–51, 2023.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[15] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[16] V. P. Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," *arXiv preprint arXiv:2012.09699*, 2020.

[17] A. Radford, "Improving language understanding with unsupervised learning," *OpenAI Res*, 2018.

[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] D. Arsovski, "GNN Book Recommender." [Online]. Available: https://github.com/arsovskii/diplomska-recommender/tree/master

Android Digital Forensics of Vulnerable Mobile Applications: Data Extraction and Analysis of Hardcoded Strings

Isiaq Bolaji Ibrahim
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, North Macedonia
isiaqibrahim@stu.khas.edu.tr

***Abstract*— This study investigates the role of Android digital forensics in addressing security vulnerabilities in mobile applications, focusing on the extraction and analysis of hardcoded strings. Hardcoded strings, often embedded directly in application code, can expose sensitive data such as API keys, credentials, and configuration details, making applications vulnerable to exploitation. The research aims to develop a systematic methodology for identifying, extracting, and analyzing these strings to assess their potential security risks.**

Keywords*— *Android digital forensics, vulnerable mobile applications, data extraction, hardcoded strings, mobile device security

I. Introduction

The rapid adoption of Android devices worldwide has led to a corresponding rise in the development and deployment of mobile applications. With Android's open-source nature and dominant market share, it has become a primary target for cyberattacks and digital forensic investigations. Android digital forensics plays a crucial role in uncovering evidence, identifying vulnerabilities, and safeguarding user data within this ecosystem.

One critical area of concern in Android application security is the presence of hardcoded strings. These strings, often embedded directly within application code, can include sensitive information such as API keys, user credentials, encryption keys, and server configurations. Developers frequently use hardcoded strings for convenience or oversight, but their presence can pose significant security risks. Attackers can exploit these strings to compromise applications, access unauthorized data, or perform malicious activities.

This study aims to address these challenges by exploring methodologies for the extraction and analysis of hardcoded strings in Android applications. By leveraging digital forensic techniques, the research seeks to enhance the understanding of these vulnerabilities and contribute to the development of more secure mobile applications.

1.1 Problem Statement

Hardcoded strings in Android applications represent a persistent security challenge that can have far-reaching consequences. These strings often include sensitive data, such as authentication tokens, API keys, encryption credentials, and server configurations, embedded directly within the application's source code. While developers may rely on hardcoded strings for convenience or lack of awareness, their inclusion poses significant risks, including unauthorized access to systems, data breaches, and exploitation by malicious actors.

Existing research on Android application security has extensively addressed broader issues such as malware detection and system vulnerabilities. However, there is a noticeable gap in studies specifically focusing on the forensic analysis of hardcoded strings. Current tools and methodologies often lack the precision or scope to systematically extract, categorize, and analyze these strings, leaving critical vulnerabilities inadequately addressed. Furthermore, many investigations fail to provide actionable insights into secure coding practices or the broader implications of such vulnerabilities for user privacy and cybersecurity.

This study seeks to bridge these gaps by developing a comprehensive framework for the digital forensic analysis of hardcoded strings in Android applications. By identifying patterns, assessing security risks, and proposing mitigation strategies, this research aims to enhance both forensic methodologies and application development practices to better safeguard user data and system integrity.

1.2 Objectives

- Analyze Hardcoded Strings: Systematically identify and examine the types and purposes of hardcoded strings embedded in Android application packages (APKs).
- Demonstrate Forensic Extraction Techniques: Develop and refine methods for extracting hardcoded strings from APK files using reverse engineering tools such as APKTool and JADX.
- Assess Security Implications: Evaluate the risks associated with hardcoded strings, including their potential to expose sensitive data and compromise application security.
- Propose Mitigation Strategies: Provide recommendations for developers to minimize the use of hardcoded strings and adopt more secure coding practices.
- Enhance Digital Forensic Practices: Contribute to the advancement of forensic methodologies for analyzing mobile application vulnerabilities, with a focus on scalability and reproducibility.

1.3 Significance of the Study

The significance of this study lies in its potential to address critical gaps in Android digital forensics and mobile application security. By focusing on hardcoded strings, this research highlights an often-overlooked aspect of application vulnerabilities that can have severe implications for cybersecurity and user privacy.

From a forensic perspective, this study provides a systematic framework for identifying and analyzing hardcoded strings, enhancing the precision and reliability of digital forensic investigations. This methodology can be integrated into existing forensic practices, aiding professionals in uncovering evidence, assessing risks, and preventing security breaches more effectively.

For developers, the findings offer actionable insights into secure coding practices. By identifying common vulnerabilities and proposing mitigation strategies, the research encourages a shift toward more secure development processes, reducing the likelihood of introducing hardcoded strings into application code.

Additionally, the study contributes to broader cybersecurity efforts by raising awareness of the risks associated with hardcoded strings. Organizations, developers, and policymakers can leverage these insights to implement better security protocols, protect user data, and maintain trust in digital platforms.

Ultimately, this research advances the field of Android digital forensics and fosters the development of more secure mobile applications, benefiting users, developers, and cybersecurity professionals alike.

2 Literature Review

2.1 Overview of Android Forensics

The field of Android forensics has evolved significantly alongside the growth of the Android operating system, now dominating the global mobile market. Android's open-source nature and customizable framework make it a popular choice for developers and users but also a primary target for malicious activities. Digital forensics for Android aims to uncover evidence, investigate breaches, and secure data through systematic analysis.

Aiman Al-Sabaawi and Ernest Foo (2019) conducted a comparative study on Android mobile forensics, evaluating manual, logical, and physical acquisition methods for retrieving data from file systems. Using a Samsung Android device (version 4.2.2), the study employed tools such as AFLogical, AccessData FTK Imager, and Autopsy. The research highlighted the complexities of Android forensic processes, particularly regarding hardware diversity and proprietary software. Data such as contacts, SMS, MMS, call logs, and media were successfully retrieved. The study emphasized the necessity of both open-source and commercial tools to ensure comprehensive forensic investigations and the importance of standardized procedures to adapt to evolving technologies.

Hussein Farooq Tayeb and Cihan Varol (2019) reviewed methodologies and tools employed in Android mobile device forensics, emphasizing the unique challenges posed by Android's open-source ecosystem and diverse hardware implementations. The study outlined the key stages of mobile forensics, including evidence identification, preservation, acquisition, examination, and presentation. It discussed the challenges associated with extracting data from volatile and non-volatile memory and highlighted tools like CelleBrite and AccessData FTK. The review underscored the significance of app-based forensics and memory imaging while advocating for standardized frameworks to address the diversity of Android devices and forensic requirements.

2.2 Hardcoded Strings in Mobile Applications

Hardcoded strings refer to textual data embedded directly into an application's source code rather than being stored in external or dynamic resources. They serve various purposes during application development, including the storage of default configurations, API keys, encryption credentials, server URLs, and error messages. While this practice can expedite development by simplifying the integration of these elements, it introduces significant security vulnerabilities.

2.3 Concept and Typical Uses

Developers often use hardcoded strings for convenience, to reduce the complexity of managing external resources, or to ensure seamless functionality during runtime. For instance:

- Authentication and API Integration: Hardcoded API keys allow applications to interact with external services without requiring additional configuration.
- Default Configuration Settings: Predefined values for app behavior are sometimes embedded in code.
- Debugging and Testing: Hardcoded credentials or test strings are used during development and testing phases.

2.4 Risks and Vulnerabilities

Despite their utility, hardcoded strings expose applications to numerous risks:

- Sensitive Data Exposure: Attackers who decompile APK files can easily access sensitive information such as passwords, API keys, and server credentials.
- Breach of User Privacy: Exposed hardcoded strings may compromise user data stored or processed by the application.
- Increased Exploitation Risk: Attackers can use the exposed information to impersonate legitimate users, bypass authentication, or execute malicious actions on associated systems.

2.5 Broader Implications

The risks posed by hardcoded strings extend beyond individual applications. Compromised keys or credentials can serve as gateways for widespread attacks, including unauthorized data access, system manipulation, and further exploitation of linked services. For users, this translates into privacy violations, identity theft, and financial losses.

While the issue of hardcoded strings has been recognized for years, its prevalence remains alarmingly high due to a combination of factors, such as:

i. Developer Oversight: A lack of awareness or inadequate security training often leads to the inclusion of sensitive hardcoded data.
ii. Absence of Code Reviews: Weak quality control mechanisms fail to identify and mitigate this vulnerability.

iii. Limited Use of Secure Alternatives: Developers frequently neglect available best practices, such as external configuration files, environment variables, or encryption methods.

The literature highlights its criticality but often lacks comprehensive solutions tailored to forensic contexts. This paper addresses these gaps by proposing a systematic forensic approach to identifying and analyzing hardcoded strings in Android applications, contributing to the fields of mobile security and digital forensics.
This section aligns with the broader scope of your research by detailing the concept, utility, and risks of hardcoded strings in Android applications. It provides a theoretical foundation to support the methodologies and findings in subsequent sections.

2.6 Existing Tools and Techniques

A range of tools and techniques has been developed to aid in the detection and analysis of hardcoded strings within Android applications. The following section reviews prominent methods currently in use:

2.6.1 Reverse Engineering Tools

- APKTool: A powerful tool for disassembling and reassembling APK files. It enables researchers to extract the underlying resource files and manifest configurations, providing insights into hardcoded elements.
- JADX: A decompiler that converts APK files into readable Java source code, making it easier to locate hardcoded strings and analyze their functionality.

2.6.2 Static Analysis Tools

- MobSF (Mobile Security Framework): Provides static analysis of APK files, highlighting potential security flaws, including the presence of hardcoded sensitive data.
- QARK (Quick Android Review Kit): Focuses on identifying security vulnerabilities in Android applications, including exposed API keys and hardcoded credentials.

Dynamic Analysis Tools

- Frida and Xposed Framework: Useful for runtime analysis, these tools can intercept function calls and observe the behavior of hardcoded strings in real-time.

2.7 Research Gaps

The forensic analysis of hardcoded strings in Android applications remains an underexplored area, despite its critical importance for both cybersecurity and user privacy. Existing studies have primarily focused on broader Android security challenges, such as malware detection, data encryption weaknesses, and system vulnerabilities. While these contributions are invaluable, they fail to address the unique risks posed by hardcoded strings.

2.7.2 Limitations of Previous Studies

i. Narrow Focus: Most research overlooks the specific security implications of hardcoded strings, concentrating instead on general application vulnerabilities.
ii. Tool Limitations: Existing tools often lack the granularity to detect and analyze obfuscated or encrypted hardcoded strings effectively.
iii. Absence of Comprehensive Frameworks: There is no systematic methodology for extracting, categorizing, and assessing the security risks associated with hardcoded strings.
iv. Limited Practical Insights: Few studies provide actionable recommendations for developers to mitigate these vulnerabilities or integrate secure coding practices.

2.8 Justification for Research

This study seeks to fill these gaps by developing a robust framework for the forensic analysis of hardcoded strings. By leveraging advanced tools, reverse engineering techniques, and contextual analysis, the research aims to:

i. Provide a systematic approach to identifying and extracting hardcoded strings.
ii. Analyze the security implications of these strings within real-world applications.
iii. Offer practical recommendations for developers and forensic practitioners to enhance mobile application security.

By addressing these gaps, the study contributes to the fields of digital forensics, mobile application development, and cybersecurity, offering insights that can improve both research and practice.

3 METHODOLOGY

The methodology is designed to systematically investigate the presence, extraction, and analysis of hardcoded strings in Android applications, ensuring a rigorous and reproducible approach.

3.1 Research Design

This study adopts a mixed-methods approach, combining both qualitative and quantitative methodologies to achieve a comprehensive analysis. The mixed-methods framework facilitates a holistic investigation, enabling the study to:

i. Qualitatively Analyze: Explore the types and purposes of hardcoded strings embedded in Android applications, including their contexts and potential security implications. This involves manual inspection and categorization of extracted data.
ii. Quantitatively Measure: Assess the prevalence and patterns of hardcoded strings across a dataset of Android applications. Statistical techniques are applied to identify trends, correlations, and the overall risk profile associated with these vulnerabilities.

The combination of these approaches ensures that the study captures both the nuanced details of specific hardcoded strings and broader trends within the dataset, providing actionable insights for both forensic practitioners and developers.

3.1.1 Tools and Techniques

The research employs a range of tools and techniques tailored for reverse engineering, data extraction, and analysis. These include:

- APKTool: For disassembling APK files and accessing resource files.
- JADX: For decompiling APKs into human-readable Java code.
- MobSF (Mobile Security Framework): To automate the static analysis of APK files and identify potential vulnerabilities.
- Custom Scripts: Developed in Python to extract and categorize hardcoded strings systematically.

3.1.2 Sampling and Dataset

The dataset comprises a diverse selection of Android applications sourced from public repositories such as Google Play Store, GitHub, and APKMirror. Applications are chosen to represent various categories, including finance, communication, and utilities, to ensure a broad analysis. The sample size includes 24 applications, providing a statistically significant basis for the findings.

Procedures

i. Data Collection: APK files are downloaded and organized based on application categories.

ii. Data Extraction: Tools like APKTool and JADX are used to reverse-engineer the applications and extract hardcoded strings.

iii. Categorization: Extracted strings are categorized based on their types (e.g., API keys, credentials, URLs) and assessed for potential vulnerabilities.

iv. Analysis: Statistical and qualitative analyses are conducted to identify trends, common vulnerabilities, and their implications for application security.

3.2 Data Collection

The data collection process for this study focuses on selecting and analyzing Android applications to extract and examine hardcoded strings. This section provides a detailed description of the steps involved, including the selection criteria, tools used, and ethical considerations.

3.2.1 Application Selection

The dataset for this research consists of Android application packages (APKs) sourced from publicly available repositories such as the Google Play Store and APKMirror. The selection process involves the following steps:

i. Categorization: Applications are chosen to represent a wide range of categories, such as finance, communication, social media, and utilities. This ensures the dataset captures diverse functionalities and potential vulnerabilities.

ii. Popularity and Ratings: Priority is given to widely used applications with significant user bases to assess real-world impacts.

iii. Permission Analysis: Applications requesting extensive permissions are specifically targeted, as these may indicate a higher likelihood of containing sensitive hardcoded strings.

iv. Exclusion Criteria: Applications with limited functionality or those primarily used for educational purposes (e.g., sample or test apps) are excluded to focus on genuine, actively used applications.

3.2.2 Reverse Engineering Process

To extract hardcoded strings, each APK undergoes a reverse engineering process using industry-standard tools:

i. Decompilation: APK files are decompiled using tools like APKTool and JADX to access underlying code and resource files.

ii. String Extraction: Extracted code is systematically searched for hardcoded strings, focusing on sensitive elements such as API keys, credentials, URLs, and encryption keys.

iii. Categorization: Identified strings are categorized based on their type, purpose, and potential security impact.

iv. Validation: The context of each string is analyzed to ensure accurate classification and identify associated risks.

3.3 Ethical Considerations

Ethical guidelines are strictly followed throughout the data collection process to ensure responsible research practices:

i. Publicly Available Applications: Only publicly accessible APK files are used, avoiding proprietary or unauthorized software.

ii. No Exploitation: Extracted data is analyzed purely for research purposes, with no attempts made to exploit vulnerabilities or compromise user data.

iii. Responsible Disclosure: Any significant security issues identified during the study are reported to developers or relevant authorities in accordance with responsible disclosure protocols.

iv. User Privacy: The study does not involve accessing or analyzing user-generated data, ensuring the privacy and security of end users are not compromised.

By adhering to these procedures and ethical standards, the data collection process ensures the reliability and integrity of the research findings while respecting legal and ethical boundaries.

4 Results

4.1 Overview of Findings

The findings of this research highlight key patterns and trends in the usage of hardcoded strings within Android applications, offering critical insights into potential security vulnerabilities.

Prevalence of Hardcoded Strings

i. High Frequency: A significant number of applications across various categories include hardcoded strings, with certain categories like finance and communication showing particularly high instances of sensitive elements.

ii. Common Types: API keys and URLs were the most frequently encountered string types, often embedded within authentication routines and data transmission functions.

4.2 Case Studies

This section highlights specific examples of vulnerable Android applications analyzed during the study, illustrating the risks posed by hardcoded strings and their broader implications for cybersecurity. By examining real-world applications, these case studies provide a practical context for understanding the security challenges associated with hardcoded elements.

Case Study 1: Finance Application with Exposed API Keys

- Application Details: A popular finance application with over 1 million downloads was found to contain hardcoded API keys in its source code.
- Vulnerability: The API keys were used to access financial data services and lacked proper obfuscation or access controls.
- Risk Analysis:

i. Potential abuse of the keys could allow unauthorized transactions or data breaches.

ii. The keys could be exploited to retrieve sensitive user financial data, leading to significant privacy violations.

- Impact: Exposing API keys not only risks service abuse but also undermines user trust and violates compliance standards such as GDPR or PCI DSS.

Case Study 2: Social Media Application with Hardcoded Credentials

Application Details: A widely used social media platform with over 5 million users was discovered to store hardcoded administrator credentials in its codebase.

Vulnerability: The credentials were accessible through reverse engineering and could provide unrestricted administrative access to the backend system.

Risk Analysis:

i. An attacker could gain control of user accounts or access sensitive data stored on the platform.

ii. The lack of secure storage for credentials increased the application's exposure to brute force and unauthorized access attacks.

• Impact: This vulnerability posed a significant threat to user privacy and platform integrity, with potential for large-scale account compromise.

Case Study 3: E-Commerce Application with Hardcoded Encryption Keys

Application Details: An e-commerce app serving over 2 million customers included hardcoded encryption keys used for securing payment transactions.

- Vulnerability:

i. The encryption keys were embedded directly in the source code without obfuscation.

ii. A predictable encryption algorithm was used, further weakening the security framework.

- Risk Analysis:

i. Exposed encryption keys could allow attackers to decrypt sensitive payment information, leading to financial fraud.

ii. The absence of dynamic key generation reduced the resilience of the application to attacks.

- Impact: The vulnerability endangered customer payment information, risking financial losses and reputational damage.

4.3 Cross-Validation with Automated Tools

Results were cross-verified using automated static analysis tools such as MobSF and manual inspection techniques. This dual approach ensures accuracy and reduces the likelihood of false positives or overlooked vulnerabilities.

4.4 Industry Benchmarks

Findings were evaluated against industry standards for secure coding practices, such as OWASP's guidelines. Discrepancies identified highlight critical areas where developers deviate from best practices, further reinforcing the study's implications.

4.5 Implications of Validation

The alignment of this study's results with prior research and industry benchmarks confirms its reliability and relevance. The validation process underscores the pervasive nature of hardcoded strings and their associated risks, emphasizing the urgency of addressing these vulnerabilities in Android applications. Furthermore, the study's reproducible methodology sets a foundation for future research and continuous improvement in the field of Android digital forensics and application security.

5. Discussion

5.1 Interpretation of Findings

The results of this research highlight critical vulnerabilities in Android applications caused by hardcoded strings, including API keys, authentication credentials, and URLs. These findings have significant implications for both Android forensics and application security:

5.2 Comparisons with Previous Research

The results of this research highlight critical vulnerabilities in Android applications caused by hardcoded strings, including API keys, authentication credentials, and URLs. These findings have significant implications for both Android forensics and application security:

5.2.1 Similarities with Prior Research:

Consistent with studies by Normaziah A. Aziz, et al., 2015, this research confirms the widespread use of hardcoded strings across various application categories. For example, the prevalence of API keys and authentication credentials reflects patterns documented in previous analyses of Android vulnerabilities.

The identified risks associated with hardcoded strings—such as exposure of sensitive information and increased susceptibility to unauthorized access—echo findings from earlier works emphasizing the critical need for secure coding practices.

5.2.2 Advancements Over Previous Studies:

Unlike many earlier studies that primarily relied on automated tools, this research integrates manual verification techniques to enhance the accuracy of string extraction and classification, reducing false positives.

The inclusion of a broader dataset covering diverse application categories provides a more comprehensive understanding of trends and risks compared to studies that focused on narrower subsets, such as financial apps or gaming platforms.

5.3 Practical Applications

The findings of this research have significant practical applications in both digital forensics and application security. Key areas where these results can be applied include:

Enhancing Forensic Methodologies:

i. The detailed process for extracting and analyzing hardcoded strings provides a reliable framework for forensic investigators. By incorporating these techniques, investigators can uncover critical evidence, such as user activity logs, API interactions, or potential attack vectors, in compromised or suspect applications.

ii. The systematic validation process introduced in this study ensures that extracted data is credible and actionable, minimizing the risk of errors in forensic reports.

5.4 Comparisons with Previous Research

The results of this research highlight critical vulnerabilities in Android applications caused by hardcoded strings, including API keys, authentication credentials, and URLs. These findings have significant implications for both Android forensics and application security:

6. Conclusion

This study has made significant contributions to the field of Android digital forensics by exploring the extraction and analysis of hardcoded strings in vulnerable mobile applications. By detailing a systematic methodology for identifying and analyzing these vulnerabilities, this research provides a practical framework for forensic investigators and application developers alike.

The findings underscore the critical role of addressing hardcoded string vulnerabilities to enhance application security and protect user privacy. Hardcoded strings, often overlooked in the development process, pose significant risks, including unauthorized data access, security breaches, and compliance violations. This study highlights the prevalence and impact of these vulnerabilities, offering actionable insights into their detection and mitigation.

In conclusion, the study not only advances the field of Android digital forensics but also serves as a call to action for stakeholders to prioritize security in application design and forensic practices. Continued research and collaboration in this area are essential to addressing the challenges posed by emerging threats and ensuring a safer mobile environment for all users.

References

[1] Acharya, S., Rawat, U., & Bhatnagar, R. (2022). A comprehensive review of Android security: Threats, vulnerabilities, malware detection, and analysis. Security and Communication Networks.

[2] Almehmadi, T., & Batarfi, O. (2019). Impact of android phone rooting on user data integrity in mobile forensics. 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 1–6.

[3] Almuqren, A., Alsuwaelim, H., Rahman, M. M. H., & Ibrahim, A. A. (2024). A systematic literature review on digital forensic investigation on Android devices. Procedia Computer Science, 235, 1332–1352. https://doi.org/10.1016/j.procs.2024.04.126

[4] Al-Dhaqm, A., Razak, S. A., Ikuesan, R. A., Kebande, V. R., & Siddique, K. (2020). A review of mobile forensic investigation process models. IEEE Access: Practical Innovations, Open Solutions, 8, 173359–173375.

[5] Al-Sabaawi, A., & Foo, E. (2019). A comparison study of Android mobile forensics for retrieving file systems. International Journal of Computer Science and Security (IJCSS), 13(4), 148–166.

[6] Aziz, N. A., Mokti, F., & Nadhar, M. (2015). Mobile device forensics: Extracting and analysing data from an Android-based smartphone. Proceedings of the IEEE CyberSec Conference. https://doi.org/10.1109/CyberSec.2015.32

[7] Casey, E. (2018). Handbook of digital forensics of multimedia data and devices. John Wiley & Sons.

[8] Ceballos Delgado, A. A., Glisson, W. B., Grispos, G., & Choo, K.-K. R. (2022). FADE : A forensic image generator for android device education. WIREs Forensic Science, 4(2).

[9] Fernando, V. (2021). Cyber forensics tools: A review on mechanism and emerging challenges. 2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 1–7.

[10] Hermawan, T., Suryanto, Y., Alief, F., & Roselina, L. (2020). Android forensic tools analysis for unsend chat on social media. 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 233–238.

[11] Mirza, Mohammad M., Salamh, F. E., & Karabiyik, U. (2020). An android case study on technical anti-forensic challenges of WhatsApp application. 2020 8th International Symposium on Digital Forensics and Security (ISDFS), 1–6.

[12] Mykhaylova, O., Fedynyshyn, T., & Platonenko, A. (2024). [Hardcoded credentials in Android apps: Service exposure and category-based vulnerability analysis] (https://ceur-ws.org/Vol-3550/). CEUR Workshop Proceedings.

[13] National Institute of Justice. (2017). Electronic crime scene investigation: A guide for first responders.

[14] National Institute of Standards and Technology. (2016). NIST Special Publication 800-86: Guide to Integrating Forensic Techniques into Incident Response.

[15] National Institute of Standards and Technology (NIST). (2020). Special publication 800-204. Guide to forensic analysis of mobile devices.

[16] Riadi, I., Umar, R., & Firdonsyah, A. (2017). Identification Of Digital Evidence On Android's Blackberry Messenger Using NIST Mobile Forensic Method. International Journal of Computer Science and Information Security (IJCSIS), 15(5), 3-8.

[17] Tayeb, H. F., & Varol, C. (2019). Android mobile device forensics: A review. Proceedings of the IEEE International Symposium on Digital Forensics and Security. https://doi.org/10.1109/ISDFS.2019.8757469

[18] Zurier, S. (2023, March 2). Financial apps tested from Google Play Store leaked sensitive API data under testing conditions. Approov Mobile Threat Lab. Retrieved from https://info.approov.io & https://www.scworld.com/news/financial-apps-google-play-leaked-data

Analysis and Comparison of Super-Resolution Models

Boris Manev
Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, North Macedonia
boris.manev@students.finki.ukim.mk

Dime Gjorgjiev
Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, North Macedonia
dime.gjorgjiev@students.finki.ukim.mk

Sasho Gramatikov
Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, North Macedonia
sasho.gramatikov@finki.ukim.mk

***Abstract*—This paper presents a comprehensive analysis of three state-of-the-art models for image super-resolution: Real-ESRGAN, HAT, and SwinIR. We evaluate these models on diverse low-resolution media types, including degraded cartoons, extremely blurry natural scenes and textured surfaces. Using both objective metrics (PIQE scores) and visual assessment, we compare their performance across restoration quality, artifact suppression, and computational efficiency. Our findings reveal that while each model offers distinct advantages, HAT consistently outperforms the others with superior detail preservation and minimal presence of artifacts. This paper highlights the importance of considering both quantitative measurements and subjective human perception when evaluating super-resolution techniques, which provide valuable insights for researchers and the public in selecting appropriate models for their specific image restoration tasks.**

Index Terms—Super-resolution, image restoration, deep learning, transformers, Real-ESRGAN, HAT, SwinIR, GAN, CNN.

I. Introduction

Image super-resolution is a very challenging task in computer vision, aiming to enhance the resolution of low-quality images to achieve high quality outputs. This process has a wide range of real-world applications, including restoring old media, medical imaging, satellite image analysis, forensic investigations and even in video game visuals [1]. Super-resolution models use advanced algorithms and deep learning to reconstruct missing details and improve overall image quality, making it indispensable for both scientific research and everyday use cases. With the rise of popularity of Machine Learning, several advancements have been developed to create these new models with remarkable upscaling capabilities. These techniques have evolved from traditional interpolation-based techniques to more advanced deep learning methods that can accurately predict missing pixels and produce a high-resolution image. This paper focuses on comparing three prominent super-resolution models: Real-ESRGAN (Enhanced Super-Resolution Generative Adversarial Network), HAT (Hybrid Attention Transformer, and SwinIR (Swin Image Restoration). Each model represents a different approach to the super-resolution problem.

II. Evolution of Image Restoration Techniques

Image super-resolution has evolved significantly over the years, from traditional interpolation techniques to advanced deep learning methods. This evolution has led to remarkable improvements in image quality and computational efficiency. In this section, we'll explore the journey from simple interpolation methods to sophisticated neural network architectures that have revolutionized the field of image super-resolution.

A. Traditional Image Interpolation Techniques

The simplest form of interpolation is Nearest Neighbor Interpolation. This technique assigns the value of the closest pixel to each neighboring pixel. Because of its simplicity the algorithm is extremely computationally efficient, but the cost is a blurry output and with very visible square-shaped pixel patterns. Another method is Bilinear Interpolation. This method considers the closest 2x2 neighbors of a selected pixel and averages their color value. This produces a smoother result than the Nearest Neighbor Interpolation, but it lacks sharpness. Bicubic Interpolation was developed as a direct improvement over bilinear interpolation, considering the closest 4x4 neighborhood around the target pixel. It uses cubic polynomials to calculate pixel values, where pixels closer to the chosen pixel hold higher weight in calculation [2]. This technique became the standard for a long while, being used in everything from NASA satellite imagery, medical imaging, to becoming a feature in all commercial image editing software.

B. Deep Learning Revolution

SRCNN (Super-Resolution Convolutional Neural Network) is the first ever deep learning solution for super-resolution. The SRCNN architecture consists of three layers: patch-extraction, non-linear mapping, and reconstruction [3]. It's simple compared to its successors, but at the time, it achieved impressive results compared to the primitive interpolation methods. ESCPN (Efficient Sub-Pixel CNN) introduces sub-pixel convolution to replace deconvolutional layers, solving an existing problem with previous models, where checkerboard artifacts would often appear [4]. This concept not only improved quality, but also significantly reduced computational requirements. TTSR (Learning Texture Transformer Network for Image Super-Resolution) is one of the first models to use the transformer [5] architecture for image super-resolution tasks. It uses the low-resolution images as queries and

reference images as keys in a transformer architecture [6]. This type of setup allows the model to learn the relationships between features in both images. The model looks for part of the reference image that matches what's needed in the low-resolution image. These revolutionary transformers are used in two of the models we will be analyzing: HAT and SwinIR.
These models and architectures each came with revolutionary innovations and paved the way for models with advanced Transformer and GAN-based techniques.

III. Super-Resolution Models

There are numerous powerful super-resolution models, each approaching the image restoration problem through different methodologies and technologies. We have selected three architecturally diverse models to evaluate their capabilities in addressing this challenging problem.

A. Real-ESRGAN

Like ESRGAN before it (and SRGAN before that) Real-ESRGAN uses a GAN [7] (Generative Adversarial Network) architecture for real-world image degradation [8]. SRGAN pioneered perceptual loss for super-resolution [9], while ESRGAN introduced Residual-in-Residual Dense Blocks without batch normalization [10]. Real-ESRGAN further advances the technology by incorporating sinc filters to replicate ringing and overshoot artifacts (visual distortions where ringing appears as wavy ripples or oscillations extending from sharp edges in an image). It is trained on a diverse dataset of synthetically downscaled images and excels at enhancing low-quality images affected by blur, noise and compression artifacts.

B. SwinIR

SwinIR uses Transformer based architecture for super-resolution. Unlike the adversarial approach of GANs, SwinIR uses self-attention mechanisms and shifted windows to capture long-range dependencies in images. The architecture consists of shallow feature extraction, deep feature extraction using Swin Transformer blocks, and high-quality image reconstruction [11]. This approach allows SwinIR to achieve excellent performance with fewer parameters than traditional CNNs, the Transformer-based method makes it particularly skilled at handling image restoration tasks with lower computational costs.

C. HAT (specifically Real Hat Gan srx4)

A hybrid model that combines the strengths of both Transformers and GANs for image restoration tasks. It utilizes channel attention for capturing global statistics and windows-based self-attention for local feature fitting, and an overlapping cross-attention module to improve interactions between neighboring features [12]. It activates more input pixels for reconstruction, giving it performance improvements in upscaling.

IV. Methodology

For our comparative analysis, we created a dataset of original, unprocessed images to evaluate the restoration capabilities of these models. The dataset includes:

- Textures of different materials, with diverse patterns and surfaces.
- Frames from animated media, specifically 80s and 90s cartoons and video games, with vibrant color pallets.
- Extremely low-resolution images, designed to test the model's ability to enhance degraded visuals.

The test images in this paper were specifically chosen to represent diverse real-world scenarios where super-resolution with the goal of image restoration might likely be applied. The cartoon image represents the restoration challenge of animated media, where preserving artistic style is as important as enhancing details. The extremely blurry tree image tests the model's ability to recover structures from severely degraded inputs. The brick wall image tests the ability to reconstruct textures accurately.

For this evaluation, we will use the Perception-based Image Quality Evaluator (PIQE) metric to assess the quality of the restored images. PIQE is a no-reference metric, that evaluates upscaled images without requiring access to the original image. PIQE analyzes spatially active regions of an image (areas with high variability) to identify distortions like noise, blocking artifacts, and texture degradations [13].

Lower PIQE scores indicate better perceptual quality, while higher scores indicate severe distortions. We specifically refrain from using references metrics such as the Structural Similarity Index (SSIM) when processing artifact-heavy source images (see Case B) because SSIM would punish legitimate detail recovery in those cases, while PIQE gives an objective quality score.

V. Results

All model evaluations were conducted on a standardized hardware environment comprising an Intel Core i5-8250U processor and 12GB of RAM, establishing a consistent baseline for qualitative comparisons of output fidelity. The visual results from the comparison of the models are shown in Figure 1. We will be analyzing the images one by one visually, and then empirically, and come to a conclusion of which one of them performs the best in three of our scenarios.

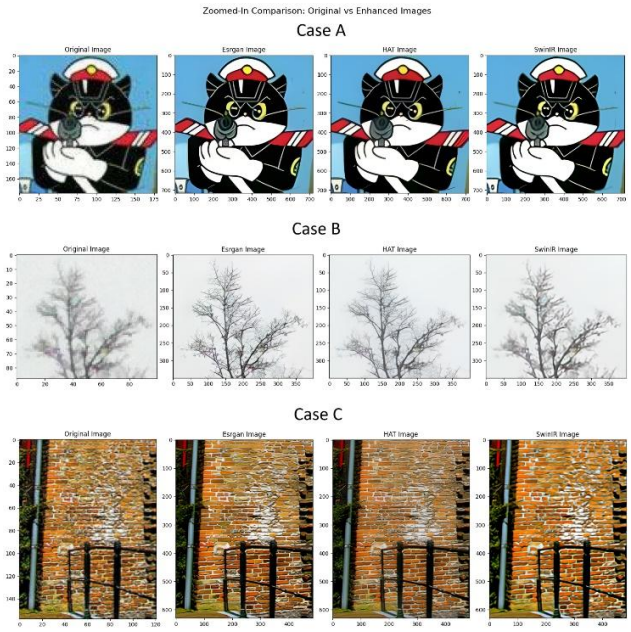


Figure 1. Comparison of the models in three different cases

A. *Cartoon Image*

The original image is heavily degraded with obvious artifacts and blurring.
Real-ESRGAN restores sharpness but introduces oversaturation in the colors and artifacts (the bottom of the pistol, whiskers of the cat).
HAT provides a more balanced restoration, preserving details with minimal saturation, but pertains the artifacts (but they're less apparent than in Real-ESRGAN).
SwinIR keeps the artifacts to a minimum, and improves clarity of the original, but struggles with detail recovery compared to HAT and Real-ESRGAN.

B. *Tree Image*

The original image is extremely blurry and lacks detail.
Real-ESRGAN enhances sharpness, but when looking closer artifacts and distortions are noticeable especially in areas dense with branches.
HAT restores finer details in the branches and maintains a high level of detail.
SwinIR slightly improves the quality of the original and struggles with the details.

C. *Brick Wall Image*

The original shows severe texture degradation and blur.
Real-ESRGAN restores the textures but introduces oversaturation and artifacts in high-frequency areas.
HAT excels at recovering brick textures while maintaining accurate colors and contras.
SwinIR produces good results in high-frequency areas, keeping the accuracy of the brick layout, but sacrificing the quality of the detail and the original color of the image.

Figure 2 shows a graph of the average PIQE scores for the three models (scaled from 0-1 to 0-100)

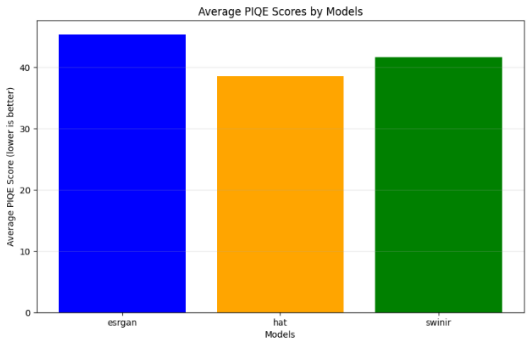


Figure 2. PIQE Graph

Real-ESRGAN has the highest score (~40), suggesting it produces lower-quality restorations compared to the other models.
HAT achieves the lowest score (~30), indicating superior restoration quality among the 3 models.
SwinIR performs slightly worse than HAT (~40), but better than Real-ESRGAN.

VI. Discussion

This evaluation highlights a fundamental challenge in image restoration research – the subjective nature of quality assessment metrics. While metrics like PIQE aim to provide objective distortion metrics, they cannot fully capture human preferences for the aesthetic appeal of images. Empirical data is not entirely enough to give conclusions in an area such as art. For example, SwinIR's second place PIQE ranking conflicts with its underwhelming performance in tasks like texture reconstruction (ex. Case C Brick wall), or upscaling very noisy images (ex. Case B Tree). This inconsistency reveals the limitations of using only quantitative metrics for results that cannot be quantified.
Image restoration and upscaling exist in a unique research place, where numerous metrics and hands-on human judgement must work together. Even sophisticated algorithms like PIQE can't accurately judge contextual factors – like animated works where the original artistic intent and originality of the art style is just as, if not more important than higher-resolution or technical accuracy.

VII. Conclusion

In this paper, a comparative analysis of super-resolution models (Real-ESRGAN, SwinIR, and HAT) was conducted. Our work highlights the strengths and weaknesses of each model in various image restoration tasks. HAT emerges as the most robust model, achieving the lowest PIQE scores (indication superior perceptual quality) and being the best in visual restoration tasks such as texture recovery, artifact suppression, and color fidelity. Its innovative architecture – combining both channel attention and window-base self-attention schemes makes it incredibly effective for real world, severely degraded images. SwinIR demonstrates competitive performance in certain cases like animated media, and has the best

computational efficiency, but struggles with detail, especially when upscaling very low-resolution inputs. Real-ESRGAN, while effective at sharpening low-resolution inputs, introduces oversaturation and artifacts that degrade the perceptual quality, as reflected in its higher PIQE scores.

In this analysis, out of the 3 models, HAT proved to be the most effective both in preserving the originality and details of the image, and its perceived quality, owing to its innovative hybrid architecture.

References

[1] P. Milanfar, (Ed) "Super-Resolution Imaging," 1st Edition, CRC Press, 2011. DOI: 10.1201/9781439819319.

[2] R. Keys, "Cubic convolution interpolation for digital image processing," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 6, pp. 1153-1160, December 1981. DOI: 10.1109/TASSP.1981.1163711.
eprint: https://ncorr.com/download/publications/keysbicubic.pdfhttps://ieeexplore.ieee.org/document/1163711.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295-307, Feb. 2016. DOI: 10.1109/TPAMI.2015.2439281.
eprint: http://arxiv.org/pdf/1501.00092.pdf.

[4] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1874-1883. DOI: 10.1109/CVPR.2016.207.
eprint: https://arxiv.org/pdf/1609.05158.pdf.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), 2017, pp. 5998-6008. DOI: 10.48550/arXiv.1706.03762. eprint: https://arxiv.org/pdf/1706.03762.pdf.

[6] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning Texture Transformer Network for Image Super-Resolution," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5791-5800. DOI: 10.48550/arXiv.2006.04139.
eprint: https://arxiv.org/pdf/2006.04139.pdf.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," in Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), 2014, pp. 2672-2680. DOI: 10.5555/2969033.2969125.
eprint: http://papers.neurips.cc/paper/5423-generative-adversarial-nets.pdf.

[8] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 1905-1914. DOI: 10.48550/arXiv.2107.10833. eprint: https://arxiv.org/pdf/2107.10833.

[9] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4681-4690. DOI: 10.1109/CVPR.2017.19. eprint: https://openaccess.thecvf.com/content_cvpr_2017/papers/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.pdf.

[10] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018. DOI: 10.1007/978-3-030-11021-5_5. eprint: https://openaccess.thecvf.com/content_ECCVW_2018/papers/11133/Wang_ESRGAN_Enhanced_Super-Resolution_Generative_Adversarial_Networks_ECCVW_2018_paper.pdf.

[11] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 1833-1844. DOI: 10.48550/arXiv.2108.10257.
eprint: https://arxiv.org/pdf/2108.10257.

[12] X. Chen, X. Wang, W. Zhang, X. Kong, J. Zhou, Y. Qiao, and C. Dong, "HAT: Hybrid Attention Transformer for Image Restoration," arXiv preprint arXiv:2309.05239, 2023. DOI: 10.48550/arXiv.2309.05239.
eprint: https://arxiv.org/pdf/2309.05239.

[13] MathWorks. PIQE - Perception based Image Quality Evaluator (PIQE) no-reference image quality score, last accessed 2025/03/28. [Online]. Available: https://www.mathworks.com/help/images/ref/piqe.html

Navigating the Data Space Landscape: Concepts, Applications, and Future Directions

Bojana Marojevikj
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
bojana.marojevikj@finki.ukim.mk

Riste Stojanov
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
riste.stojanov@finki.ukim.mk

***Abstract*—This paper explores the evolving landscape of data spaces, focusing on key concepts, practical applications, and emerging future directions. It begins by introducing the foundational principles that underpin data space architectures, emphasizing their importance in facilitating secure and efficient data exchange. The core design principles and essential building blocks that form the backbone of data-space systems are then examined. Several real-world implementations are presented, showcasing how data spaces are applied across various industries to address challenges such as data sovereignty, interoperability, and trust. The paper concludes by discussing future directions, proposing that leveraging semantic data models can significantly enhance interoperability and data integration within data spaces. Furthermore, it suggests exploring the implementation of SPARQL as a sophisticated authorization layer to improve security and granular control over data access. This research provides a comprehensive understanding of the current state of data spaces and aims to guide future advancements in this rapidly evolving field by highlighting the potential of semantic data and SPARQL-based authorization.**

***Index Terms*—Data spaces, International Data Spaces Association, Eclipse Dataspace Connector, Reference Architecture Model, Gaia-X, SPARQL**

I. Introduction

In an era characterized by exponential data growth and the increasing need for seamless data exchange, the concept of data spaces has emerged as a potential solution. The purpose of data spaces is to break down data silos and enable trusted data sharing, fostering innovation across ecosystems. They create a common ground where multiple stakeholders, such as businesses, governments, research institutions, and even individual users, can securely share and collaborate on data while maintaining control over who can access what. This approach not only boosts efficiency but also unlocks new opportunities for collaboration and value creation. For example, in healthcare, a data space can allow hospitals, research laboratories, and pharmaceutical companies to share anonymized patient data for research purposes. This could accelerate drug discovery, improve personalized medicine, and improve patient outcomes, while complying with strict privacy regulations such as GDPR [1]. Data spaces rely on technologies such as identity management, access control, and data usage policies to ensure secure interactions. Initiatives like Catena-X [2] in Europe illustrate how these ecosystems work in practice, promoting transparency, interoperability, and digital sovereignty. This paper will explore the core concepts of data spaces, highlight real-world applications, discuss challenges, and outline future directions. By understanding how data spaces function and their transformative potential, we can better navigate the evolving digital landscape and unlock new frontiers of innovation.

II. Core Concepts

The concept of data spaces emerged around 20 years ago within the field of computer science as an alternative to data sharing through restrictive bilateral agreements between organizations [3]. Unlike centralized methods, such as data consolidation hubs, data spaces do not require physically merging data into a single location. Instead, data remains stored at its original source. Additionally, a unified database schema is not necessary for integrating data from multiple sources. Instead, integration occurs at the semantic level through shared vocabularies. This flexibility allows for data redundancies and the coexistence of different datasets. Data spaces can also overlap or nest within each other, allowing participants to belong to multiple data spaces at the same time [4]. One of the first real implementations of the concept of data space emerged in 2015 with the International Data Space (IDS) initiative [5].

A. Design Principles

The core design principles that guide the development and operation of data spaces, ensuring that they are secure, fair, and sustainable are the following:

- **Data Sovereignty:** Data sovereignty refers to the right of individuals and organizations to independently control their economic data assets. This is a core principle in data spaces, empowering participants to: (1) manage, process, and secure their data more effectively, and (2) retain control over data access when sharing with others. [6]
- **Level Playing Field for Data Sharing and Exchange:** Creating a fair environment within data spaces ensures that all participants, including new entrants, can compete based on data quality and service offerings rather than data volume. This fairness is supported by promoting

collaboration over competition, achieved through careful design and ongoing maintenance of the underlying infrastructure. [6]

- **Decentralized Soft Infrastructure:** The infrastructure for data spaces is supposed to be decentralized and composed of interoperable systems adhering to shared functional, technical, operational, and legal standards. Although invisible to participants, this "soft infrastructure" should handle interoperability, security, privacy, and trust through advanced identity and consent management. It should also support data monetization, while remaining technology neutral to allow participants flexibility in choosing their tools and platforms. [6]
- **Public-Private Governance:** Effective governance is crucial for building and sustaining data spaces. It requires representation from businesses, individuals, governments, and technology partners to balance public and private interests. Existing regulations (such as GDPR [1] and eIDAS [7]) provide a foundation, and new policies are being developed to further support data spaces. The public sector will initially drive adoption through funding and participation, with long-term success driven by network effects and community growth. [6]

B. Building blocks

The architectural landscape of data spaces is characterized by several recurring patterns that reflect the core principles of the concept. A prominent pattern is that of a decentralized architecture, where data remains under the control of the originating producers, and access to these data is granted in a carefully managed and secure manner. This approach stands in contrast to centralized data repositories, such as data lakes. Another significant pattern is the federated architecture, which involves the interconnection of multiple data spaces based on a foundation of shared policies and rules. This federation enables data sharing across diverse industries, companies, and entities, creating a network of interconnected data stores rather than a single monolithic repository. Additionally, some initiatives, such as the International Data Spaces Association (IDSA) [5], propose the IDSA Reference Architecture Model (RAM) [8]), a conceptual model which serves as a blueprint with the purpose of providing common architectural principles and guidelines for organizations to create or join a Data Space. It may be considered as the current leading architecture reference for Data Space initiatives and can be seen as a set of technical components within the broader context of building blocks as shown in Figure 1.

Several essential technical components are integral to the functioning of data spaces. As illustrated in Figure 2, data sharing in a data space involves several key technical components working in coordination. **Connectors** serve as the critical technical interface, facilitating data exchange by implementing and enforcing the policies established within the data space. Connectors manage the negotiation of contracts between data providers and consumers and ensure adherence to the usage rules defined by participants. They also enable secure data

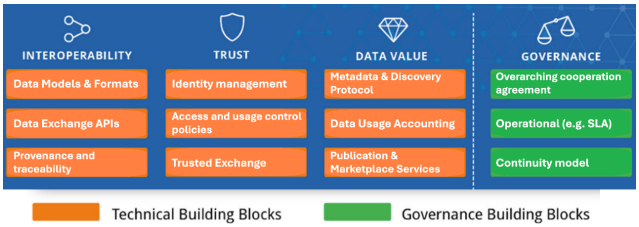


Fig. 1. An overview of technical and governance building blocks for data spaces [9]

transfer through a variety of protocols and application programming interfaces (APIs), such as Java Database Connectivity (JDBC) [10], Amazon Simple Storage Service (Amazon S3) [11], and Hypertext Transfer Protocol Secure (HTTPS) [12]. Examples of connector technologies include Eclipse Dataspace Connector (EDC) [13], FIWARE TRUE Connector [14], and Simpl [15]. The effectiveness of data spaces depends on these connectors, which provide the technical means for secure and policy-compliant data sharing across distributed systems.

Data catalogs and **discovery services** play a vital role in enabling participants to register and locate available data assets and services within the data space ecosystem. These components store metadata pertaining to data offerings, including crucial details such as usage policies and access conditions. They also support functionalities for searching and browsing the array of available data sources. Efficient data discovery is crucial for data spaces, enabling participants to find required data while adhering to providers' policies.

Identity and Access Management (IAM) systems are fundamental to ensuring a secure data sharing environment by managing identities and authentication process. This often involves the use of digital identities and verifiable credentials to establish trust among participants and facilitate secure transactions. Robust IAM capabilities are essential for building confidence and ensuring that only authorized entities can access and utilize data within defined policy parameters.

Policy enforcement mechanisms are critical for managing the agreements between data providers and consumers prior to any data exchange. These mechanisms ensure that agreed-upon data usage policies are strictly enforced, thereby guaranteeing compliance with the conditions stipulated by data owners. This may involve sophisticated contract negotiation processes and the application of policy languages, such as the Open Digital Rights Language (ODRL) [16]. Automation of policy enforcement is vital for maintaining data sovereignty and ensuring that data are used in accordance with the agreed terms, thus minimizing the need for manual oversight.

Finally, shared services encompass a range of functionalities that support the operation and governance of the data space. These can include services for participant registration and validation, certification processes, monitoring of data space activities, and overall data space management functions. The management and provision of these shared services can be

centralized or decentralized, depending on the specific requirements and design of the data space. Shared services are crucial for ensuring a level playing field for all participants and maintaining the overall integrity and functionality of the data space ecosystem.

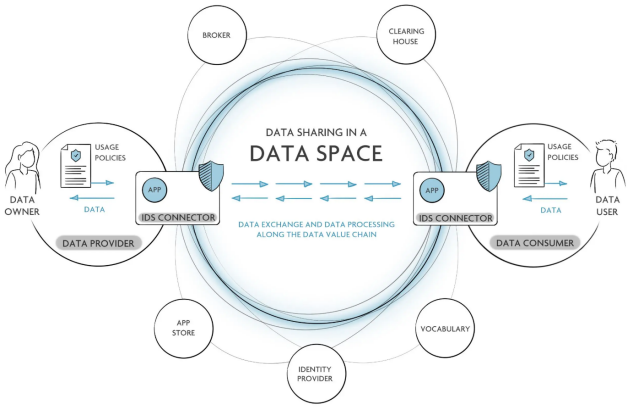


Fig. 2. Data sharing in a data space [17]

III. Existing Technical Developments and Initiatives

The data space landscape is being actively shaped by several key organizations and initiatives that are driving standardization, providing frameworks, and fostering collaboration. The International Data Spaces Association (IDSA) [5] is an association dedicated to creating standards for data sharing within data spaces, thereby empowering participants with complete control over their data assets. IDSA Reference Architecture Model (RAM) [8] is a crucial framework for the design and implementation of data spaces. The association's core objectives revolve around promoting data sovereignty and ensuring interoperability across diverse data ecosystems, underpinned by governance principles emphasizing accountability, transparency, fairness, and responsibility.

Gaia-X [18] represents a significant European initiative aimed at developing a federated and secure data infrastructure grounded in European values of transparency, openness, data protection and security. With the main goal of ensuring European digital sovereignty and fostering the digital economy, Gaia-X focuses on building a network of trust among industrial partners. It provides a comprehensive blueprint for the creation of secure, transparent, and interoperable data infrastructures.

The Data Spaces Support Centre (DSSC) [19] plays a coordinating and supportive role in the development of data spaces throughout Europe, with funding from the European Commission. DSSC is responsible for developing the Data Space Blueprint [20], which outlines essential building blocks and design principles for data spaces. The center provides a variety of tools and support services to address the organizational and technical requirements of establishing and operating data spaces. Its primary aim is to establish common requirements and promote best practices to accelerate the creation of sovereign data spaces within the European digital landscape.

The Data Spaces Business Alliance (DSBA) [21] represents a collaborative effort between IDSA [5], the Big Data Value Association (BDVA) [22], Gaia-X [18], and the FIWARE Foundation [23]. The primary focus of this alliance is on achieving technical convergence among existing data space architectures and models, with the goal of ensuring interoperability and portability across different data space implementations. DSBA represents a concerted effort to harmonize the diverse initiatives in the data space domain, fostering greater cohesion and interoperability within the ecosystem.

Several key open-source projects and platforms are also contributing significantly to the technical foundation of data spaces. The Eclipse Dataspace Components (EDC) [24] project provides an open source framework for building secure and globally scalable data-sharing services. EDC offers a set of customizable components for constructing control planes, data planes, decentralized identity systems, and federated data catalogs. It is built on the specifications of the Gaia-X Trust Framework [18] and the IDSA Dataspace protocol [25] and is utilized by initiatives such as Eona-X [26] and Catena-X [2]. EDC serves as a fundamental open source resource, providing the essential building blocks for implementing data spaces based on established standards and frameworks.

FIWARE [23] offers a range of open source components that can be used to build data spaces. In particular, the FIWARE TRUE Connector offers a dedicated specification to enable secure and efficient data sharing within the IDS ecosystem. FIWARE's contribution lies in providing reusable open-source tools that facilitate the development of interoperable data space solutions.

Simpl [15] is an open source, smart and secure middleware platform initiated by the European Commission to facilitate the creation of EU data spaces and enhance interoperability between them. Simpl aims to streamline data sharing across various data spaces and enable data scaling from the edge to the cloud. Its initial deployment is focused on key public sector domains such as healthcare, public procurement, and mobility.

IV. Real-World Implementations

The potential of data spaces is being realized in various industries through concrete implementations. In the logistics industry, the Smart Freight Centre (SFC) Exchange Network [27] is a collaborative network built as a data space on AWS. Its primary goal is to promote transparency and decarbonization in transport chains by facilitating the exchange and reporting of emissions data from logistics.

The automotive industry has seen the development of Catena-X [2], a prominent data space driven by more than 60 companies. This initiative aims to address challenges and opportunities in areas such as traceability, sustainability, the circular economy, and efficient supply chains throughout the automotive value chain. Catena-X heavily relies on the Eclipse Dataspace Connector (EDC) as a core technology for enabling secure and sovereign data exchange among its participants.

The healthcare sector is another area where data spaces are making significant strides. The European Health Data Space (EHDS) [28] aims to create a secure and interoperable environment for the exchange of health data, to accelerate medical research and improve healthcare delivery. Similarly, Dataspace4Health [28], with contributions from NTT DATA Luxembourg, provides a secure environment for health data exchange, enabling new medical research while protecting patient privacy.

In the realm of mobility and smart cities, the Mobility Data Space (MDS) [28] serves as a platform for aggregating, managing, and sharing mobility-related data from various sources. This facilitates data collaboration for purposes such as traffic management, urban planning, and transportation optimization. Another data space initiative in this sector is the Flanders Smart Data Space [29] which aims to establish a comprehensive and interoperable data infrastructure in the region of Flanders (Belgium). This data space leverages a combination of key technologies to achieve its objectives. The Eclipse Dataspace Connector (EDC) is utilized to provide secure and sovereign data exchange capabilities among participants. Complementing this, Semantic Web technologies are employed to ensure semantic interoperability across diverse datasets. The core technology for data dissemination within the space is Linked Data Event Streams (LDES), which enables near real-time access and synchronization of information for data consumers.

Another initiative is the Green Deal Data Space (GDDS) [30] which is designed to facilitate the secure and interoperable exchange of high-quality data between sectors to support sustainability and resilience goals. It integrates diverse datasets related to climate change, biodiversity, zero pollution, circular economy, smart mobility, and environmental compliance. Built on FAIR principles [31], the GDDS aims to overcome fragmented data systems by connecting public, private, and citizen-generated data within a trusted digital ecosystem.

In the energy sector, the development of data spaces is being supported through various EU-funded initiatives. The Energy Data Space is a key focus of the ETIP SNET policy framework [32], aiming to enable secure, standardized, and interoperable data exchange across energy systems and actors. Coordinated by the BRIDGE initiative [33], efforts in this domain have led to significant advancements in cross-dataspace interoperability. Notable testbeds include the OMEGA-X [34] and ENERSHARE [35] data spaces, which have demonstrated practical scenarios for interoperable and sovereign data sharing in energy applications. These initiatives not only address technical challenges but also contribute to the establishment of a federated ecosystem for energy data exchange aligned with European digital and energy policy goals [36].

V. Future Directions

The current landscape of data spaces, while promising, faces several critical challenges that threaten its widespread adoption and the realization of a truly interconnected data economy. One of the primary challenges lies in achieving seamless interoperability and semantic understanding across diverse data sources. Although existing standards and protocols provide a foundation, a robust semantic layer is crucial for enabling machines to interpret and process data accurately. This necessitates the development of comprehensive ontologies and knowledge graphs that can capture the context and meaning of data across various domains. Furthermore, ensuring secure and granular access control within data spaces remains a significant hurdle. Current authorization mechanisms often lack the flexibility and precision needed to manage complex data sharing scenarios. To address this, future directions should explore the integration of advanced access control techniques, such as the SPARQL authorization guard. This approach aims to create a more intelligent and secure data environment where:

- Data are annotated with semantic metadata, enabling automated discovery, integration, and reasoning. This will facilitate cross-domain data exchange and create a richer, more contextualized data landscape.
- A SPARQL authorization guard will enable fine-grained control over data access based on semantic queries and policies. This will allow data providers to define precise rules for who can access what data and under what conditions, enhancing security and trust.
- The combination of semantic interoperability and robust access control mechanisms will contribute to a more dynamic data economy. This enables data providers to monetize their assets, while allowing data consumers to securely access relevant information.

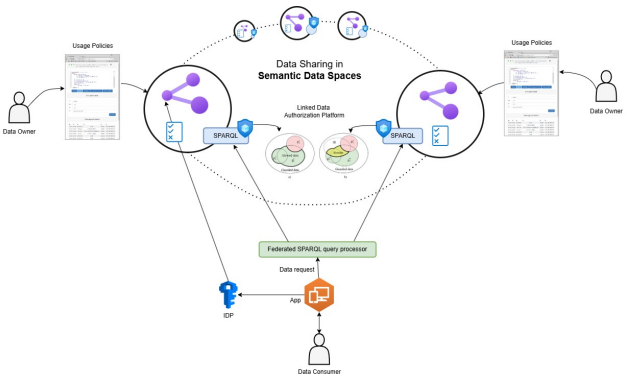


Fig. 3. Semantic data space with an authorization layer

VI. Conclusion

This research paper has explored the complex landscape of data spaces, exploring their key definitions, principles, wide-ranging applications, and potential future developments. Data spaces represent a paradigm shift in how data is managed and shared, moving away from traditional centralized systems toward decentralized, collaborative ecosystems. The key principles of data sovereignty, interoperability, trust, security, and governance underpin this novel approach, addressing critical

concerns associated with data exchange in an increasingly interconnected world.

The applications of data spaces are vast and varied, spanning industries from logistics and automotive to healthcare and public services. Real-world implementations demonstrate their potential to facilitate data integration, enhance collaboration, strengthen governance, and power data marketplaces, ultimately unlocking significant value for individuals, organizations, and society. The benefits are multifold, ranging from increased control over personal data and improved public services to fostering innovation and driving economic growth.

However, the journey towards widespread adoption of data spaces is not without its challenges. Technical complexities, data quality issues, integration hurdles, and concerns about privacy and security must be carefully addressed. The lack of standardization remains a significant challenge, highlighting the need for collaborative efforts to establish common frameworks and protocols.

Looking ahead, the future of data spaces is set to evolve with a greater emphasis on semantic interoperability, granular access control, and secure data exchange. The development of comprehensive ontologies and knowledge graphs will enhance the ability to interpret and integrate data across domains, fostering a more interconnected ecosystem. At the same time, fine-grained access control mechanisms, such as SPARQL authorization guards, will enable precise and policy-driven data sharing, strengthening security and trust. As data spaces continue to mature, ongoing research and collaboration among stakeholders will be essential in overcoming existing challenges and unlocking their full potential for a more efficient and dynamic data economy.

References

[1] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A practical guide, 1st ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[2] Catena-X, "Catena-X: The Automotive Network." https://catena-x.net/en/1. Accessed: 2025-03-15.

[3] M. Franklin, A. Halevy, and D. Maier, "From databases to dataspaces: a new abstraction for information management," *ACM Sigmod Record*, vol. 34, no. 4, pp. 27–33, 2005.

[4] B. Otto, "The evolution of data spaces," in *Designing data spaces: The ecosystem approach to competitive advantage*, pp. 3–15, Springer International Publishing Cham, 2022.

[5] I. D. S. Association, "International data spaces." https://internationaldataspaces.org/. Accessed: 2025-03-13.

[6] L. Nagel, J. J. Hierro, E. Perea, D. Lycklama, C. Mertens, A.-S. Taillandier, M. Marques, J. Gelhaar, A. Marguglio, U. Ahle, *et al.*, "Design principles for data spaces: position paper," tech. rep., E. ON Energy Research Center, 2021.

[7] A. Sharif, M. Ranzi, R. Carbone, G. Sciarretta, F. A. Marino, and S. Ranise, "The eidas regulation: a survey of technological trends for european electronic identity schemes," *Applied Sciences*, vol. 12, no. 24, p. 12679, 2022.

[8] I. D. S. Association, "Ids ram 4.0." https://internationaldataspaces.org/, 2022. Accessed: 2025-03-13.

[9] L. Nagel and D. Lycklama, "How to build, run, and govern data spaces," in *Designing data spaces: The ecosystem approach to competitive advantage*, pp. 17–28, Springer International Publishing Cham, 2022.

[10] E. Sciore and E. Sciore, "Jdbc," *Database Design and Implementation: Second Edition*, pp. 15–47, 2020.

[11] Amazon Web Services, "Amazon S3 - Cloud Object Storage." https://aws.amazon.com/s3/. Accessed: 2025-03-22.

[12] Z. Durumeric, J. Kasten, M. Bailey, and J. A. Halderman, "Analysis of the https certificate ecosystem," in *Proceedings of the 2013 conference on Internet measurement conference*, pp. 291–304, 2013.

[13] Eclipse Foundation, "Eclipse EDC Connector." https://github.com/eclipse-edc/Connector. Accessed: 2025-03-22.

[14] FIWARE Foundation, "FIWARE True Connector." https://fiware-true-connector.readthedocs.io/en/latest/. Accessed: 2025-03-22.

[15] European Commission, "Simpl: Cloud-to-edge federations empowering EU data spaces." https://digital-strategy.ec.europa.eu/en/policies/simpl. Accessed: 2025-03-22.

[16] R. Ianella, "Open digital rights language (odrl)," *Open Content Licensing: Cultivating the Creative Commons*, 2007.

[17] International Data Spaces Association, "IDSA Infographic: Data Sharing in a Data Space." https://internationaldataspaces.org/wp-content/uploads/IDSA-Infographic-Data-Sharing-in-a-Data-Space.pdf. Accessed: 2025-03-30.

[18] H. Tardieu, "Role of gaia-x in the european data space ecosystem," in *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*, pp. 41–59, Springer International Publishing Cham, 2022.

[19] Data Spaces Support Center, "Data Spaces Support Center." https://dssc.eu/. Accessed: 2025-03-15.

[20] Data Spaces Support Center, "Data Spaces Blueprint." https://dssc.eu/space/BPE/179175433/, 2024. Accessed: 2025-03-15.

[21] Data Spaces Business Alliance, "Data Spaces Business Alliance." https://data-spaces-business-alliance.eu/. Accessed: 2025-03-15.

[22] Big Data Value Association, "Big Data Value Association." https://bdva.eu/. Accessed: 2025-03-15.

[23] F. Foundation, "Fiware foundation." https://www.fiware.org/foundation/. Accessed: 2025-03-13.

[24] Eclipse Foundation, "Eclipse Data Components (EDC)." https://projects.eclipse.org/projects/technology.edc. Accessed: 2025-03-22.

[25] International Data Spaces Association, "IDS Knowledge Base - Dataspace Protocol." https://docs.internationaldataspaces.org/ids-knowledgebase/dataspace-protoco. Accessed: 2025-03-22.

[26] EONA-X, "EONA-X: The Tourism, Mobility and Logistics Data Space." https://eona-x.eu/. Accessed: 2025-03-22.

[27] Amazon Web Services, "Examples of data spaces built on top of AWS services." https://docs.aws.amazon.com/prescriptive-guidance/latest/strategy-building-data-spaces/data-spaces-on-aws-services.html, 2024. Accessed: 2025-03-15.

[28] NTT Data, "The Power of Data Spaces." https://benelux.nttdata.com/insights/whitepapers/the-power-of-data-spaces, 2024. Accessed: 2025-03-15.

[29] A. De Craene, Z. Vanlishout, and P. Colpaert, "Semantic and technically interoperable data exchange in the flanders smart data space,"

[30] V. Lush, L. Bastin, K. Otsu, and J. Masó, "Assessing fairness of citizen science data in the context of the green deal data space," *International Journal of Digital Earth*, vol. 17, no. 1, p. 2344587, 2024.

[31] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[32] E. SNET, "Energy data space – policy paper." https://horizoneuropencpportal.eu/sites/default/files/2024-05/etip-snet-energy-data-space-policy-paper-2023.pdf, 2023. Accessed: 2025-05-09.

[33] E. Commission, "Bridge initiative: Coordinating energy research and innovation projects across the eu," 2021. Accessed: 2025-05-09.

[34] OMEGA-X Consortium, "Omega-x project." https://omega-x.eu/. Accessed: 2025-05-09.

[35] ENERSHARE Consortium, "Enershare project." https://enershare.eu/. Accessed: 2025-05-09.

[36] A. Kung, "Omega-x standardisation workshop – presentation slides." https://omega-x.eu/wp-content/uploads/2025/04/Omega-X-Standardisation-workshop-Antonio-Kung.pdf, 2025. Accessed: 2025-05-09.

Time-Series Heart Rate Variability with CUDA

A. Shekerov, D. Mileski, M. Gusev
Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, Skopje, North Macedonia
E-mail: andrej.shekerov@students.finki.ukim.mk, {dimitar.mileski, marjan.gushev}@finki.ukim.mk

***Abstract*—This paper focuses on the computation of Time-Series Heart Rate Variability (HRV) metrics using the CUDA parallel computing platform. HRV represents an essential set of metrics in modern medicine and cardiology, finding use in patient diagnosis and treatment. Multiple preprocessing and filtering steps are required on patients' annotated electrocardiogram (ECG) data to eliminate noise and other unwanted distortions. From the cleaned data, metrics are calculated for overlapping time intervals, akin to a sliding window, for multiple configurations of width and offset, requiring considerable computational effort to generate results. We present a CUDA implementation of an HRV algorithm for short-term windows, capable of processing both single-day and multi-day patient recordings. An existing C++ implementation of the same algorithm is used for performance comparison, and both implementations are tested on a set of preprocessed patient databases to measure the execution time. A discussion of the benefits and drawbacks of utilizing CUDA follows the presentation of the results. The findings show that the CUDA implementation only slightly outperforms C++ for large datasets, with a maximal speedup of 1.12 times, and that C++ is a better and more flexible option for smaller datasets.**

***Index Terms*—heart rate variability, electrocardiogram, C++, GPU, CUDA**

I. Introduction

Heart Rate Variability (HRV) measures the fluctuations around the mean heart rate, representing changes in time intervals between consecutive heartbeats. These metrics are influenced by multiple factors, including posture, respiratory rate, and the absence of physiological activity, among others, and they provide valuable insights regarding the sympathetic and parasympathetic functions of the autonomic nervous system [1].

Heart Rate Variability showcases the heart's ability to adapt and respond to new stimuli quickly. These metrics evaluate the overall cardiac health and function of the autonomic nervous system. HRV can indicate the risk of sudden cardiac death in postinfarction and diabetic patients [2]. A population-based study by Dekker et al. concludes that for middle-aged men and women, elevated heart rate and low HRV predict increased mortality rates [3].

This study focuses on developing a CUDA algorithm for calculating key time-domain HRV metrics and relevant metrics for annotated patient electrocardiogram (ECG) recordings at different sliding window time intervals. These time intervals, or windows, have a specific width and offset, starting from the beginning of an ECG recording. As there is a significant amount of data overlap and calculations for each window, this process can take considerable time. Generating all short-term metrics for the Gluco23 database (647 MB of annotated ECG data [4]) takes about 17.6 minutes with a fast but sequential C++ implementation. This implementation serves as our base for execution-time measurement as we attempt to leverage the parallelization potential of the calculation process via a CUDA implementation.

Related work is presented in Section II. Section III elaborates on the implementation and experiment methodology. Results in Section IV are discussed in Section V. Section VI concludes the paper and provides future work directions.

II. Related work

Graphics Processing Unit (GPU) processing is widely utilized in research and development, enabling efficient processing and analysis of extensive datasets. Modern medicine relies on patient data gathered from specialized medical devices to provide accurate treatment globally, and methods are continuously advancing to utilize this data effectively. HRV is a crucial metric in cardiology, and research conducted by Wang et al. focuses on developing and optimizing a parallel algorithm to calculate HRV in the time domain. The authors utilize the PhysioNet open-source project for data and demonstration tools, and their parallel reduction algorithm achieves a 38-fold speedup compared to conventional serial methods [5]. However, the authors do not calculate time-series HRV metrics for sliding window intervals; instead, they focus on the entirety of patients' ECG recordings. Patients in the Gluco23 database have between 1951 and 107134 valid windows, with an average of 33,894 (excluding windows that are filtered out due to insufficient heartbeat data), each requiring its own set of metric calculations. The authors' positive results are a strong motivation to attempt GPU parallelization with our specific HRV method implementation.

Many data processing tasks in modern medicine can be tackled more efficiently using parallel programming paradigms. Konstantinidis et al. investigate the benefits of utilizing a GPU in biomedical signal processing by developing an algorithm that significantly reduces the execution time compared to a standard sequential C algorithm. Dealing with memory management issues through special design techniques enabled them to achieve a 29-fold speedup [6]. Saha et al. analyze parallel methods for computing the Correlation Dimension, a resource-intensive process on multichannel data sets that utilizes complex signal processing algorithms. GPUs and multi-core processing nodes are used in parallel to accelerate the algorithm, leveraging data from the MIT-BIH Arrhythmia Database. Multicore Processing nodes have an 18-fold increase in performance, while GPU clusters have a performance

increase ranging from 3 to 50 times [7]. In the domain of pupillometry, Nasim et al. analyze the performance gains of parallel algorithms. Complex algorithms, such as the Hough transform, used in pupil measurement, make the process computationally expensive. A parallel data distribution framework was implemented to run on multicore CPUs and manycore GPUs, achieving 1.86x and 3.56x speedups, respectively [8].

III. Methods

This section describes the HRV calculation methodology used, the implementation specifics, and details regarding the experiments employed. We present and analyze the following implementations of the HRV algorithm:

1) *Sequential C++* implementation, which calculates the metrics for each window iteratively and
2) *Parallel CUDA* implementation that uses CUDA by calculating windows in parallel, where a different block in the GPU calculates each window's metrics.

A. Heart Rate Variability Method

The HRV calculation methodology originates from [9] and is further developed in [10]. The preprocessing and filtering steps described in the original methodology have not been parallelized for this study since they run in linear time and are highly dependent on the data.

Originating from patients' annotated electrocardiogram recordings, heartbeats and their corresponding timestamp can be classified into multiple classes: Normal (*N*), Atrial (*A*), SupraVentricular (*S*), Ventricular (*V*), Fusion of Normal and Ventricular (*F*), Paced (*P*), and unclassifiable beats (*Q*). As the heartbeat class can significantly impact the resulting HRV metrics, filtering steps are taken to ensure that the time-domain HRV metrics are derived only from beats of class *N*.

The time (in seconds) between two consecutive heartbeats is denoted as an RR interval. If these heartbeats are both of Normal (*N*) class, then the time between these two heartbeats is called an NN interval. The filtering process removes all RR intervals that are not between consecutive Normal beats and returns a collection of uninterrupted and noise-free segments of NN intervals - *clean segments* (Fig 1).

Given a list of clean segments and their corresponding timestamps, we can begin extracting HRV metrics. For a single patient, these metrics are calculated for specific sliding-window time intervals with varying widths (the duration of the analyzed window) and offsets (the difference in time between consecutive windows). A valid window contains multiple clean segments, so corresponding metrics have to be calculated as the average (*A*), standard deviation (*S*), and combined (*C*) variants. A complete list of metrics used for our CUDA implementation is given in Table I.

B. Implementation

This subsection describes the behavior of both the reference C++ implementation and the newly developed CUDA implementation. Both begin by reading the filtered NN interval data for a specific patient, which has been preprocessed [10].

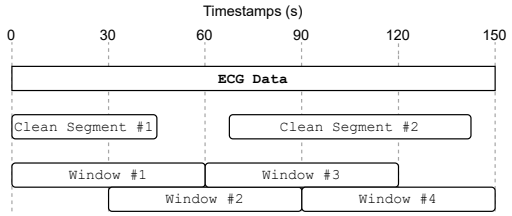


Fig. 1. ECG Data & Clean Segments

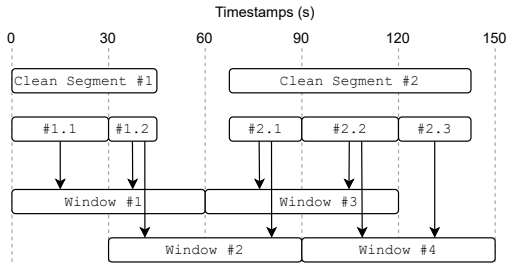


Fig. 2. Subsegments' contributions to Windows

The entire ECG array can be split into disjoint 'time intervals' at the window's predefined offset (Fig. 2). Each time interval is defined this way, and if it contains at least one beat, it is guaranteed to include the first beat for some window, and the same applies to the last beat. Following a linear preprocessing step where we identify the first and last beats for each time interval, this method enables a constant-time lookup for the indexes of any window's first and last beats, thereby determining the range of beats contained within that window.

The $window_report$ function takes a pointer to two floating-point number arrays (*R_sec* & *NN_intervals*) containing both the timestamp and value for each NN interval, as well as information regarding the window configuration (window's index, width, offset, and other parameters). Using the lookup strategy outlined above, the index of the first and last beat for the specific window is gathered, and all values in this range are stored in a newly allocated array. In the C++ implementation, this is a dynamically allocated array; in the CUDA implementation, the array resides within the block's shared memory. Calculations for all HRV metrics are now performed on the values in this array, and the results are returned once all metrics have been computed.

Both implementations utilize the $window_report$ function to generate the metrics for a specific window given its index. Calling this function for multiple window indexes simultaneously introduces parallelism. The CUDA grid is configured to create 500 blocks per kernel call, each block containing one thread, resulting in the simultaneous calculation of up to 500 consecutive windows. The gathered results are written to disk by both implementations afterward.

TABLE I
HRV METRICS FOR A SPECIFIC WINDOW [9]

Metric	Description
nNN	Number of valid intervals
nSeg	Number of clean segments
C_SDNN	Standard deviation of intervals
A_SDNN	Average of SDNN across clean segments
S_SDNN	Standard deviation of SDNN across clean segments
C_RMSSD_NN	Square root of the mean of the squares of differences between adjacent intervals
A_RMSSD_NN	Average of RMSSD across clean segments
S_RMSSD_NN	Standard deviation of RMSSD across clean segments
pNNX	Percentage of differences between adjacent intervals greater than X

TABLE II
SPECIFICATION OF WINDOW CONFIGURATIONS FOR SHORT-TERM HRV

Configuration	Width (s)	Offset (s)
1	30	30
2	60	60
3	120	60
4	240	60
5	300	60

C. Experiments

Both implementations calculate HRV for 16 preprocessed benchmark databases with five different short-term window configurations (Table II). The result is measured by running each database for each implementation five times and noting down the average execution time of the entire program. The host testing machine has the following configuration: Intel Core i7 CPU at 2.67 GHz, 11 GB system memory, and a GM200 GeForce GTX Titan X GPU.

The benchmark datasets (some originating from PhysioNet [11]) are given below:

1) AHA Database (ahadb) [12]
2) CU Ventricular Tachyarrhythmia Database (cudb) [13]
3) European ST-T Database (edb) [14]
4) St Petersburg INCART 12-lead Arrhythmia Database (incartdb) [15]
5) Long Term AF Database (ltafdb) [16]
6) MIT-BIH Long-Term ECG Database (ltdb) [17]
7) Long Term ST Database (ltstdb) [18]
8) MIT-BIH Arrhythmia Database (mitdb) [19]
9) MIT-BIH Normal Sinus Rhythm Database (nsrdb) [20]
10) MIT-BIH Noise Stress Test Database (nstdb) [21]
11) QT Database (qtdb) [22]
12) MIT-BIH ST Change Database (stdb) [23]
13) MIT-BIH Supraventricular Arrhythmia Database (svdb) [24]
14) Gluco Database 2017 (glucodb) [25]
15) Dusseldorf Clinic 2023 (dussdb) [4]
16) Gluco Database 2023 (gly23db) [4]

IV. RESULTS

Table III presents the results with information for each testing database, including the file size (in megabytes), reference execution time (in seconds), CUDA implementation's execution time (in seconds), and achieved speedup. The achieved speedup for each database, sorted in ascending order of size, is also showcased in Fig. 3.

TABLE III
EXECUTION TIME COMPARISON

database	size	cpp	cuda	speedup
cudb	0.4	0.184	0.441	0.41
nstdb	0.5	0.235	0.42	0.55
stdb	1.6	0.709	0.932	0.76
qtdb	1.9	0.831	1.173	0.70
mitdb	2.1	0.91	1.181	0.77
incartdb	3.4	1.491	1.76	0.84
ahadb	3.5	1.542	1.86	0.82
svdb	3.6	1.588	1.943	0.81
ltdb	13.0	5.689	5.703	0.99
edb	18.0	7.809	7.477	1.04
nsrdb	40.0	17.513	16.395	1.06
ltafdb	165.0	72.032	69.229	1.04
ltstdb	199.0	87.472	79.863	1.09
glucodb	329.0	150.375	140.102	1.07
dussdb	813.0	368.897	343.909	1.07
gly23db	1100.0	484.28	432.181	1.12

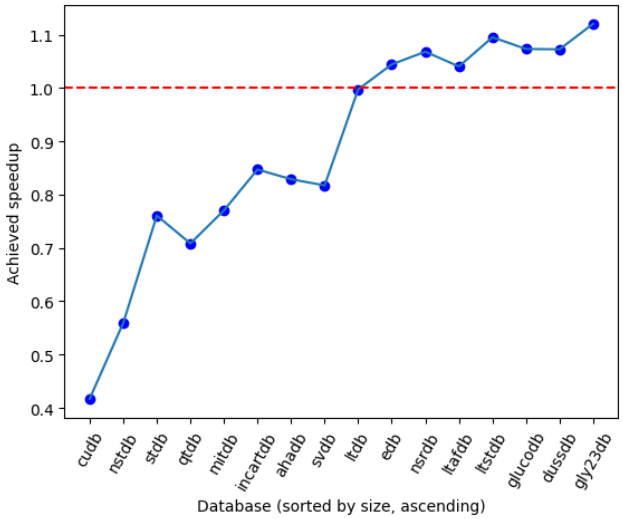


Fig. 3. Achieved speedup for each database.

V. Discussion

The results show our C++ implementation outperforms the CUDA implementation for most smaller databases, while it is slightly the opposite for larger datasets ($>$ 15MB) by a very slim margin. Even on much larger databases (glucodb, dussdb, gly23db) with multi-day patient recordings, the CUDA implementation does not achieve a significant speedup over C++.

The achieved speedup is low, and data transfer may be the cause of the performance bottleneck. The data transfer overhead is not worth the benefit of parallelization for smaller databases. For the more extensive databases, a slight improvement is evident in the findings. However, the data transfer speed to and from host memory during the initialization phase limits the potential of the parallel solution. In most practical applications, this slight edge in execution time is not worth the effort required to run the algorithm on a GPU device.

VI. Conclusion

Using our current CUDA implementation, we conclude that the benefits of GPU for calculating HRV metrics are not worth the increased cost and complexity, with only marginal efficacy improvements. By avoiding costly data transfer, the C++ implementation remains a more reliable solution. Additionally, C++ offers methods for creating shared-memory concurrency and parallelization, which can further improve execution time without relying on GPU parallelization methods.

Future work includes improving the implementation of CUDA by utilizing shared memory more efficiently. Devising an algorithm that cleverly calculates HRV metrics for common subsegments across windows via a sliding window technique using CUDA could potentially lead to improvements [26]. Alternatively, utilizing C++ shared-memory parallelization methods via OpenMP could lead to a reduction in execution time.

References

[1] C. M. van Ravenswaaij-Arts, L. A. Kollee, J. C. Hopman, G. B. Stoelinga, and H. P. van Geijn, "Heart rate variability," *Annals of internal medicine*, vol. 118, no. 6, pp. 436–447, 1993.

[2] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: a review," *Medical and biological engineering and computing*, vol. 44, pp. 1031–1051, 2006.

[3] J. M. Dekker, R. S. Crow, A. R. Folsom, P. J. Hannan, D. Liao, C. A. Swenne, and E. G. Schouten, "Low heart rate variability in a 2-minute rhythm strip predicts risk of coronary heart disease and mortality from several causes: the aric study," *Circulation*, vol. 102, no. 11, pp. 1239–1244, 2000.

[4] A. Angjelevska, M. Gusev, and S. Gjorgjieva, "Classification of hemoglobin a1c from long and extra-long term heart rate variability," in *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE, 2024, pp. 1187–1192.

[5] J. Wang, W. Chen, and G. Hou, "Parallel computing method for hrv time-domain based on gpu," in *Algorithms and Architectures for Parallel Processing: 15th International Conference, ICA3PP 2015, Zhangjiajie, China, November 18-20, 2015, Proceedings, Part II 15*. Springer, 2015, pp. 434–443.

[6] E. I. Konstantinidis, C. A. Frantzidis, L. Tzimkas, C. Pappas, and P. D. Bamidis, "Accelerating biomedical signal processing algorithms with parallel programming on graphic processor units," in *2009 9th International Conference on Information Technology and Applications in Biomedicine*. IEEE, 2009, pp. 1–4.

[7] D. P. Saha and A. R. Desai, "Performance analysis of computing multichannel correlation dimension (d2) on multicore system and gpu," *Researchgate. net*, 2012.

[8] A. Nasim, A. Maqsood, and T. Saeed, "Multicore and gpu based pupillometry using parabolic and elliptic hough transform," *International Journal of Mechanical Engineering and Robotics Research*, vol. 6, no. 5, 2017.

[9] M. Gushev, "System and method for blood glucose monitoring based on heart rate variability," Jun. 20 2024, uS Patent App. 18/279,113.

[10] A. Shekerov, A. Angjelevska, and M. Gusev, "Efficient time-series heart rate variability metrics in c++," Tech. Rep. 0202, 2025.

[11] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[12] "Aha database sample excluded record," Jun. 6 2003. [Online]. Available: https://physionet.org/content/ahadb/1.0.0/

[13] F. Nolle, F. Badura, J. Catlett, R. Bowser, and M. Sketch, "Creigard, a new concept in computerized arrhythmia monitoring systems," *Computers in Cardiology*, vol. 13, no. 1, pp. 515–518, 1986.

[14] A. Taddei, G. Distante, M. Emdin, P. Pisani, G. Moody, C. Zeelenberg, and C. Marchesi, "The european st-t database: standard for evaluating systems for the analysis of st-t changes in ambulatory electrocardiography," *European heart journal*, vol. 13, no. 9, pp. 1164–1172, 1992.

[15] E. Yakushenko, "St petersburg incart 12-lead arrhythmia database," May 1 2008. [Online]. Available: https://physionet.org/content/incartdb/1.0.0/

[16] S. Petrutiu, A. V. Sahakian, and S. Swiryn, "Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans," *Europace*, vol. 9, no. 7, pp. 466–470, 2007.

[17] G. Moody, "Mit-bih long-term ecg database," Aug. 3 1999. [Online]. Available: https://physionet.org/content/ltdb/1.0.0/

[18] F. Jager, A. Taddei, G. B. Moody, M. Emdin, G. Antolič, R. Dorn, A. Smrdel, C. Marchesi, and R. G. Mark, "Long-term st database: a reference for the development and evaluation of automated ischaemia detectors and for the study of the dynamics of myocardial ischaemia," *Medical and Biological Engineering and Computing*, vol. 41, pp. 172–182, 2003.

[19] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE engineering in medicine and biology magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[20] G. Moody, "Mit-bih normal sinus rhythm database," Aug. 3 1999. [Online]. Available: https://physionet.org/content/nsrdb/1.0.0/

[21] G. B. Moody, W. Muldrow, and R. G. Mark, "A noise stress test for arrhythmia detectors," *Computers in cardiology*, vol. 11, no. 3, pp. 381–384, 1984.

[22] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg," in *Computers in cardiology 1997*. IEEE, 1997, pp. 673–676.

[23] P. Albrecht, "St segment characterization for long term automated ecg analysis," Ph.D. dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering . . . , 1983.

[24] S. D. Greenwald, R. S. Patil, and R. G. Mark, *Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information*. IEEE, 1990.

[25] A. Stojmenski, M. Gusev, I. Chorbev, S. Tudjarski, L. Poposka, and M. Vavlukis, "Age and gender impact on heart rate variability towards noninvasive glucose measurement," *Sensors*, vol. 23, no. 21, p. 8697, 2023.

[26] G. Temelkov and M. Gusev, "Leveraging dataframe-based operations for calculation of heart rate variability," in *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE, 2024, pp. 1193–1198.

Farm and Animal Intelligent Recommendation System - FAIRS

Zoran Petrovski
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius"
Skopje, Republic of N.Macedonia
petrovski91@gmail.com

Ana Madevska Bogdanova
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius"
Skopje, Republic of N.Macedonia
ana.madevska.bogdanova@finki.ukim.mk

***Abstract*—The Farm & Animal Intelligent Recommendation System (FAIRS) is a machine learning-based decision support system that aims to help farmers and veterinarians make health-conscious decisions about the movement of livestock. This paper describes the architecture, methods, and experimental evaluation of FAIRS, which combines health data, engineered features, and predictive models that support informed animal movement authorization to recommend actionable interventions. We evaluated various classifiers (Logistic Regression, Random Forest, XGBoost, LightGBM) and data balancing techniques (SMOTE, under-sampling) on imbalanced datasets to optimize accuracy and recall in detecting at-risk animals. The results show that FAIRS is most effective, particularly when using the combination of between XGBoost and SMOTE.**

***Index Terms*—Machine learning-based decision support system, classifiers, data balancing techniques, detection of at-risk animals**

I. Introduction

Modern agricultural operations adopt data-based technologies to improve productivity levels and sustainability practices. Precision agriculture has made substantial advances in crop management, but livestock health management needs complete decision support systems [8], [11], [12]. Farmers, together with veterinarians (and/or institutions), must handle animal movement authorizations between farms and marketplaces/slaughterhouses with great caution to protect animal welfare and safety. Incorrect choices lead to false outbreaks of disease and economic losses, compromised food safety, and negative impacts on animal welfare [13]. Smart farming technologies have improved decision making processes for crop management through their advanced analytics and sensor-based monitoring [8], [11]. The current system for monitoring livestock health remains unorganized because it relies on multiple separate tools and human interpretation of data. We propose the Farm & Animal Intelligent Recommendation System (FAIRS), which operates as a single intelligent platform that addresses this challenge. The FAIRS system combines veterinary expertise with AI analytics and multiple sources of health data to generate accurate and timely recommendations and risk assessments [19]. The National Animal Identification and Traceability System (NAITS) operates independently because its initial deployment occurred in Georgia under the supervision of the Food and Agriculture Organization (FAO). FAIRS primarily addresses the following:

- Real-time risk detection in animal health and history data;
- Actionable intervention recommendations for detected health issues;
- A detailed evaluation of animal safety to prevent biosecurity risks while maintaining regulatory compliance.

II. Background

The NAITS registers contain all existing data on animal identification together with animal movements, livestock market exposure, health records, and vaccination events [8]. The FAIRS system determines animal movement safety by evaluating temperature conditions, animal activity levels, disease symptoms, recent illnesses, and exposure to livestock [19], [23]. The NAITS system operates as a centralized data storage system instead of acting as an active decision support tool despite its fundamental importance. The manual interpretation of raw data by farmers and veterinarians for animal movement decisions results in errors due to cognitive overload and incomplete data and misinterpretation of subtle health indicators [6], [19]. The Farm & Animal Intelligent Recommendation System (FAIRS) was created to address this information gap: whether an animal is safe to be moved. The FAIRS system receives the input directly from NAITS and is processing historical and real-time data through machine learning algorithms and domain-specific expertise. The system analyzes health parameters including temperature anomalies, severity of symptoms, history of illness, and exposure risks to provide interpretable recommendations on animal movement safety. This intelligent system converts NAITS from a basic registry into a dynamic decision support framework that enhances biosecurity and improves both the accuracy of decisions and operational efficiency in livestock management [19], [23].

III. Related work

The adoption of decision support systems (DSS) and predictive analytics in precision agriculture continues to grow because it improves operational efficiency and sustainability. The development of tools that match crop monitoring and yield optimization for livestock health and welfare remains restricted despite notable progress in these areas [8], [11],

[12]. According to Wolfert et al. [8] Big data continues to expand its role in smart farming, but the implementation of these technologies in animal health management systems remains underdeveloped. Several studies have investigated machine learning applications for the detection of animal diseases, but most of these implementations focus on diseases while operating independently and need specific sensors and a customized infrastructure for deployment [12], [19]. The current fragmented system prevents large-scale implementation and practical use in animal welfare programs. National animal identification systems such as NAITS function mainly as passive data storage systems which collect identification and movement data and veterinary event records, but do not provide predictive analytics or decision-making tools [8]. The lack of predictive and decision support capabilities in these systems restricts their effectiveness during urgent situations such as outbreaks of diseases and high-risk animal movements.

IV. Methodology

A. Data Collection and Pre-processing

Multiple farms provide comprehensive data collection that includes animal health indicators and behavioral patterns, as well as vaccination history and market exposure and logistics. The data quality management process involves strict cleaning procedures to address inconsistencies and outliers and inaccuracies. The numeric variables animal temperature and activity levels receive mean imputation to replace missing data points. The mode is used to impute missing values in categorical variables that include symptoms and vaccination status, and recent illnesses. The preparation of machine learning models requires one-hot encoding of categorical features to avoid unintended ordinal relationships between categories. The standardization process through scaling methods is applied to numerical data to create uniformity that minimizes training bias [1], [6], [12], [13].

B. Feature engineering

Feature engineering is critical in extracting meaningful patterns from raw data. Key engineered features include:

- Risk Score: The composite metric uses multiple critical health indicators including body temperature, severity of symptoms, history of illness, and interactions with the livestock market to provide a more detailed representation of an animal's risk profile [23].
- Interaction Features: The interaction terms include temperature times severity of symptoms, temperature times recent illness status, and activity level times temperature to detect complex non-linear relationships between individual variables. Machine learning models gain improved predictive capabilities for critical health outcomes through these interactions that individual characteristics cannot detect on their own [7],[24].

C. Data balancing

Real-world livestock datasets typically exhibit a significant imbalance, with healthy animal records vastly outnumbering those representing at-risk or unhealthy animals. This imbalance can significantly degrade the predictive precision for the minority class (animals at risk) [1], [16]. To address this issue, FAIRS employs several balance strategies, such as SMOTE [1], [17], and Random Under sampling [16]. The FAIRS system requires class imbalance solutions because real-world livestock health datasets contain significantly more records of healthy animals than at-risk animals. SMOTE enhances the training data representation of at-risk animals by creating new examples that avoid duplicate entries to maintain data diversity for better model generalization. The random undersampling method simultaneously reduces the majority class examples by choosing a smaller representative subset. The combined strategies improve the model's ability to detect minority-class patterns, which leads to better recall and earlier identification of potential health risks.

D. Machine Learning Models

Four classification algorithms were selected for evaluation in the FAIRS system based on their complementary strengths in interpretability, robustness, and suitability for unbalanced health-related datasets [3], [2], [4]:

- Logistic regression was included due to its simplicity, computational efficiency, and interpretability. It provides a probabilistic framework for binary classification and allows a straightforward understanding of the influence of characteristics, which is particularly valuable in domains where decision transparency is essential, such as veterinary risk assessment [5], [6].
- Random Forest, a bagging-based ensemble learning method, was chosen for its proven capacity to handle high-dimensional data and capture complex nonlinear interactions between variables [3], [10].
- XGBoost was incorporated because of its state-of-the-art performance in classification tasks, particularly on imbalanced datasets. Its robustness and scalability make it well suited for detecting rare but significant animal health anomalies [2], [17].
- LightGBM was selected for its ability to handle large-scale data with a high training speed and lower memory consumption. Its support for native handling of categorical features and parallel learning capabilities further enhances its practicality in resource-constrained or time-sensitive decision- making environments [4], [25].

V. Experiments and Results

The experimental framework served to evaluate four classification algorithms using data balancing techniques that were specifically designed for unbalanced livestock health datasets. The research evaluated Logistic Regression and Random Forest and XGBoost and LightGBM through tests with and without balancing strategies which included both SMOTE and random under sampling. The experiments followed a uniform preprocessing and evaluation approach to maintain methodological consistency [1], [16], [17]. The experimental pipeline included the following components:

ML model	No balancing	SMOTE
Logistic regression	x	x
Random Forest	x	x
XGBoost	x	x
LightGBM	x	x

TABLE I
COMBINATIONS USED

- Dataset: All models executed their training and evaluation through a unified data set that contains animal health records together with behavioral indicators and farm management data [12], [23].
- Train-Test Split: Training and testing data received an 80/20 stratified split that maintained the original class distribution throughout both subsets for an accurate evaluation of model performance [14].
- Pre-processing: The imputation of missing numerical values used mean values, and categorical features used mode values for the imputation. The conversion of categorical variables to binary representations used one-hot encoding and numerical features received standardization to address scale differences [6].
- Feature Engineering: The expressiveness of the model improved through the implementation of multiple domain-specific features. The Risk Score calculation used indicators that included abnormal body temperature and severity of symptoms, as well as a recent history of illness and exposure to livestock market environments. The models received enhanced risk representations through engineered interaction terms that included temperature × symptom severity, temperature × recent illness and activity level × temperature [23], [24].

A. Evaluation of ML models and balancing techniques

To compare and optimize the FAIRS results, the models were combined with the data balancing techniques used. The following experimental scenarios were targeted and compared with the experiment:

B. Evaluation metrics

The experiments used various evaluation metrics to provide a thorough evaluation of the model performance because each metric measured different aspects of the classification quality.

- Accuracy measures the proportion of correct predictions in total instances to evaluate the correctness of the model. The single use of this metric proves insufficient in assessing unbalanced classification problems [7], [24].
- The inclusion of precision and recall metrics became essential because health-related applications require these metrics. Precision calculates the ratio of actual positive predictions to all positive predictions, which demonstrates the model's capability to prevent incorrect alarms. The recall metric evaluates the correct identification of actual positive cases, while false negatives require special attention because early detection of animal health risks depends on this metric [6], [14].

Model	**Data balancing**	**Accuracy**	**Precision**	**Recall**
Logistic regression	-	0.952	1.00	0.95
Logistic regression	SMOTE	0.9615	1.00	0.96
Random Forest	-	0.96	0.93	0.99
Random Forest	SMOTE	0.93	0.89	0.99
XGBOOST	-	0.9755	0.99	0.99
XGBOOST	SMOTE	0.974	0.99	0.98
LightGBM	-	0.9745	0.99	0.98
LightGBM	SMOTE	0.9755	0.99	0.99

TABLE II
EVALUATION METRICS: ACCURACY, PRECISION, RECALL - RESULT FOR EACH COMBINATION

- The F1 score provides a balanced evaluation metric through its calculation of precision and the recall harmonic mean, which helps to evaluate models when false positives and false negatives have equal costs. The metric works best for unbalanced datasets containing one class with a much higher number of instances than the other [7], [24].
- The receiver operating characteristic curve ROC-AUC provided a method to evaluate the effectiveness of the models in distinguishing between classes through different classification thresholds. A higher value of the ROC-AUC demonstrates a superior separation between healthy and at risk categories, making this metric an effective tool to evaluate overall discrimination ability [22].

C. Results

The experimental results provide a detailed comparison of the four evaluated classifiers—Logistic regression, Random Forest, XGBoost, and LightGBM—under different data balancing conditions: without balancing, with SMOTE, and with Random Undersampling. Performance was assessed using accuracy, precision, recall, F1 score, and ROC-AUC to ensure a comprehensive evaluation in both balanced and unbalanced scenarios. These results are provided in *Fig.2* The results show that tree-based ensemble models, XGBoost and LightGBM, performed better than Logistic Regression in terms of recall and general discrimination capability, especially when used with SMOTE. This improvement shows the effectiveness of synthetic minority oversampling in increasing the sensitivity of the models to the minority (at-risk) class. XGBoost had the best balance between Precision and Recall and was the best model for identifying animals at elevated health risk. Its strong F1 score and ROC-AUC values also confirm its robustness in managing class imbalance while maintaining high classification accuracy. These results emphasize the importance of selecting suitable classifiers and data balancing techniques when developing decision support systems for health-based livestock movement assessments. Table III summarizes the experimental results of different classifiers in varying data-balancing scenarios.

VI. DISCUSSION

The experimental results shed important light on the practical application of the FAIRS system as a decision support tool for health-informed livestock movement control and the

Model	Data balancing	F1	AUC
Logistic regression	-	0.97	0.9873
Logistic regression	SMOTE	0.98	0.9872
Random Forest	-	0.96	0.9959
Random Forest	SMOTE	0.94	0.9977
XGBOOST	-	0.99	0.9903
XGBOOST	SMOTE	0.99	0.9903
LightGBM	-	0.99	0.9901
LightGBM	SMOTE	0.99	0.9901

TABLE III
EVALUATION METRICS: F1-SCORE, AUC RESULT FOR EACH COMBINATION

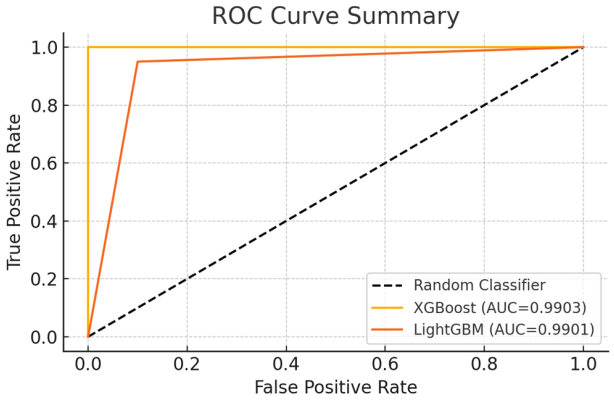


Fig. 1. ROC curve representation of the best models

methodological suitability of its implementation. Tree- based ensemble classifiers including Random Forest, XGBoost, and LightGBM showed superior performance compared to other models in all evaluation metrics, especially when handling real-world data imbalance scenarios. The models achieved high precision and recall when identifying at risk animals through ensemble learning frameworks and their capacity to model complex non-linear relationships [2], [3], [4]. XGBoost produced the most reliable results among classifiers in all balancing methods, suggesting its reliability for early disease detection needs. The model's ability to maintain stable performance under varying class distribution patterns indicates its strong generalization ability across different agricultural settings [17]. Logistic regression performed well under balanced conditions, but its linear classification failed to perform under class distribution skewness. The combination of SMOTE with logistic regression brought a significant improvement in recall while preserving precision levels. The research findings demonstrate the common relationship between model interpretability and predictive capacity that has been well documented in the literature [6], [18]. Logistic regression continues to be a valuable decision system model for transparent applications, yet it falls short when modeling complex livestock health data interactions. This study validates the essential function of data balancing approaches in veterinary epidemiology, since it requires the identification of infrequent but critical at-risk cases. The application of SMOTE improved the detection of minority classes, which led to the system detecting early warning signs of biosecurity threats and enabling more precise veterinary intervention [1], [14]. The deployment of FAIRS for real-time applications requires additional research. The system requires performance optimization to handle streaming data environments in large-scale farm networks [20]. The integration of SHAP [10] and LIME [9] explainability frameworks will give stakeholders clear explanations of system recommendations, increasing user trust and regulatory compliance, and improving accountability in livestock movement choices [13], [15].

VII. CONCLUSION

The research introduces FAIRS as a scalable and intelligent decision support system designed to tackle critical challenges in regulating livestock movement based on animal health. FAIRS demonstrates its effectiveness in identifying and interpreting high-risk animal health situations by leveraging domain-specific feature engineering, ensemble learning models, and SMOTE techniques for handling imbalanced real-world datasets. The experimental results demonstrate that the tree-based XGBoost and LightGBM classifiers, together with suitable preprocessing pipelines, produce substantial predictive performance. Both ML techniques and data preparation methods are crucial for developing reliable decision support systems that serve agricultural biosecurity needs. The stakeholders in the livestock industry gain direct value through FAIRS because the system provides data-based decision support that helps to prevent diseases early while ensuring the welfare of the animals and meeting regulatory requirements [13], [14]. The system connects the passive data collection capabilities of NAITS to the need for active veterinary operational guidance that produces interpretable results. Future research will focus on implementing FAIRS in operational settings and adding IoT health monitoring systems and XAI techniques, including SHAP and LIME to improve model interpretability and stakeholder trust [9], [10], [20]. The modular design and extensibility of FAIRS indicate its potential to become a fundamental model for next generation digital farming systems that can be applied in multiple agricultural applications [15].

REFERENCES

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.
[3] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[4] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
[5] S. Menard, Applied Logistic Regression Analysis, vol. 106. Sage, 2002.
[6] C. Molnar, Interpretable Machine Learning. Lulu Press, 2020.
[7] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
[8] S. Wolfert, L. Ge, C. Verdouw, and M. J. Bogaardt, "Big Data in Smart Farming—A Review," Agricultural Systems, vol. 153, pp. 69–80, 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.
[10] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[11] A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Boldú, "A Review on the Practice of Big Data Analysis in Agriculture," Computers and Electronics in Agriculture, vol. 143, pp. 23–37, 2017.
[12] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review," Sensors, vol. 18, no. 8, 2018.
[13] L. Klerkx, E. Jakku, and P. Labarthe, "A Review of Social Science on Digital Agriculture, Smart Farming and Agriculture 4.0," NJAS - Wageningen Journal of Life Sciences, vol. 90, 100315, 2019.
[14] H. Suresh and J. V. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning," Communications of the ACM, vol. 64, no. 11, pp. 62–71, 2021.
[15] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2021.
[16] J. Brownlee, Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-sensitive Learning, Machine Learning Mastery, 2020.
[17] A. Fernández et al., "Learning from Imbalanced Data Sets," Springer Briefs in Machine Learning, Springer, 2018.
[18] Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.
[19] M. T. McCabe and J. R. Ryan, "Machine Learning Applications in Animal Health Surveillance Systems," Veterinary Informatics Journal, vol. 35, no. 2, pp. 104–115, 2020.
[20] D. H. Tran, A. Phung, and S. Venkatesh, "Streaming Decision Trees for Big Data Analytics," in IEEE Int. Conf. Data Science and Advanced Analytics (DSAA), 2016.
[21] G. Holmes, A. Donkin, and I. H. Witten, "Weka: A Machine Learning Workbench," in Proc. Second Australia and New Zealand Conf. Intelligent Information Systems, 1994.
[22] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[23] R. Bisdorff, "Feature Engineering for Predictive Models in Animal Welfare Assessment," Journal of Agricultural Informatics, vol. 13, no. 1, pp. 35–49, 2022.
[24] D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient over F1 Score and Accuracy in Binary Classification Evaluation," BMC Genomics, vol. 21, 2020.
[25] Y. Barua, "Explainable AI for Agriculture: Challenges and Opportunities," in Proc. IEEE Int. Conf. on AgriTech, 2023.

Enhancing AI Education with a Custom Code Runner Template: A Solution for Fair Grading and Improved Feedback in Search-Based Problems

1st Martin Dinev
FCSE, UKIM
Skopje, North Macedonia
martin.dinev@finki.ukim.mk

***Abstract*—Artificial Intelligence courses often rely on search-based problem-solving exercises to teach uninformed and informed search algorithms. At FCSE Skopje, students practiced these concepts using a searching framework, but grading challenges arose due to the limitations of Moodle's Code Runner plugin. The traditional approach only compared student output to a single expected solution, leading to incorrect grading when multiple valid solutions existed. Additionally, verifying student understanding became time-consuming due to widespread code plagiarism.**

To address these issues, we developed a Python-based template for the Code Runner plugin that evaluates solutions by validating action sequences rather than static outputs. The template ensures that student-generated solutions correctly transition between states and reach a goal state while also verifying function correctness across different test cases. By implementing execution isolation between student code and the correct solution, and by providing detailed feedback, this approach enhances grading fairness, discourages cheating, and improves student learning. The template is designed for easy adaptation to various problems using the searching framework, making it a scalable and efficient tool for AI education.

***Index Terms*—Automated Grading, Searching Framework, Moodle Code Runner, Artificial Intelligence Education, Student Feedback System**

I. Introduction

Teaching the practical aspect of search algorithms in AI courses comes with unique grading challenges. At FCSE Skopje, students use a searching framework to implement uninformed and informed search strategies. While Moodle's Code Runner plugin was used to automate grading, it had serious limitations: it could only compare student output to a single expected solution. This caused frustration when students submitted correct solutions that were marked as incorrect simply because they didn't match the expected output.

Another major issue was cheating. Since students received assignments before lab sessions, many copied solutions rather than implementing them themselves. Instructors then had to spend valuable lab time verifying whether students understood their submissions instead of helping them improve their problem-solving skills.

To fix these problems, I developed a Python template for Moodle's Code Runner that grades solutions based on their action sequences rather than final outputs. This ensures that students are fairly evaluated even when they produce different but correct solutions. The template also isolates student code execution from the reference solution and provides detailed feedback to help students understand where their implementation fails. This paper explains how the template was designed, how it works, and how it improves grading fairness, student engagement, and instructor efficiency.

II. Background & Related Work

The concept of automated grading has been under development for over half a century, with the earliest systems focusing on objective question types [1]. The evolution has progressed from basic scoring mechanisms to sophisticated technology-enhanced assessment and feedback systems designed to support both teachers' grading efforts and students' learning journeys [1]. The rapid growth in student enrollments, especially in computer science and programming disciplines, has amplified the need for and reliance on automated systems like Automated Grading Systems (AGSs) [1]. These systems are increasingly seen as high-quality and sustainable solutions for managing the demands of assessing larger student cohorts [1]. Traditionally, automated grading in computer science courses has relied on one of the following approaches:

- Output-Based Grading: The most common method of automated grading compares the output of a student's code against a predefined correct output. If the results match exactly, the solution is considered correct. This method is efficient for deterministic problems, such as mathematical computations or simple programming tasks. However, it is less suitable for problems where multiple valid outputs exist, such as those involving search algorithms.
- Unit Testing and Functional Testing: Some grading systems evaluate submissions by running a set of unit tests, checking whether specific functions return the expected values for given inputs. This approach is commonly used in test-driven development (TDD) environments and is useful for verifying that individual components of a program work correctly. However, it does not always account for the overall logic and correctness of an algorithm in state-space search problems.
- Code Similarity Detection for Plagiarism Prevention: Since automated grading makes it easier for students to submit copied solutions, many systems incorporate

plagiarism detection tools such as MOSS (Measure of Software Similarity). These tools analyze structural similarities in student submissions to detect potential academic dishonesty. However, they do not address the issue of verifying whether students truly understand their solutions.

In the context of AI education at FCSE, Output-Based Grading has been utilized for immediate feedback, followed by manual review from the teaching assistant or lab tutor during lab exercises. However, this approach presented challenges for both students who could solve the problems independently and those who struggled. The former faced frustration and doubt when their correct solutions were marked as incorrect due to rigid output matching, while the latter had difficulty identifying whether their code contained real errors, often requiring extensive manual debugging without clear guidance. This was also a challenge for the TAs/tutors during lab exercises, as they had to balance helping students, grading and validating solutions within a limited time, and determining whether a student had worked independently or plagiarized the solution.

A. Existing Searching Framework and `Problem` Class Implementation

The searching framework already in use provides students with a `Problem` class, in the python programming language, that they are required to extend and use by supplying it to predefined implementations of uninformed and informed search algorithms. This class serves as the interface through which students model their specific problem scenarios. It required the implementation of five core methods:

- A **constructor**, which provides the problem with an initial state,
- `successor(state)`, which returns a dictionary mapping available actions to their resulting states,
- `actions(state)`, which typically returns the keys from the `successor(state)` dictionary,
- `result(state, action)`, which returns the resulting state for a given action, usually derived from the `successor(state)` output, and
- `goal_test(state)`, which evaluates whether a given state is a goal state.

With these methods the students have freedom to model their problems based on their understandings and their own ideas for how to represent the state in any given problem. They get to choose which searching algortihm to use. At last they are required to print the actions for the path returned by the searching algorithm. Since the action names are listed in the problem statement, this sequence of actions can be compared to the sequence of actions derived from the correct solution from the problem setters.

The environment that is used is the plugin Code Runner for Moodle. It uses Output-Based Grading but also has advanced functionalities for setting your own templates for grading the students submissions.

III. Proposed Solution

The existing `Problem` class appears to be a natural candidate for applying Unit Testing and Functional Testing as a grading method. However, since students are given the freedom to model their own representations of problem states, we cannot enforce a standardized input or output format for testing the methods in their implementations. One possible workaround would be to prescribe the structure of the state in the problem statement, but this unfortunately reveals part of the solution. Modeling the state is an important learning objective, and prescribing it would undermine the purpose of the exercise.

That said, because students are required to print a sequence of actions that lead from the initial state to the goal state, we can validate their solutions by checking whether the length of the sequence matches that of a known correct solution. Ideally, we would go further and verify whether applying these actions actually transitions through valid states and ultimately reaches a goal state. The only reliable way to perform such validation is to run the correct implementation of the `Problem` class (written by the problem setter) using the sequence of actions produced by the student.

To accomplish this within Moodle's Code Runner plugin, we developed a custom template written in Python. The template executes the student's submission to retrieve the printed output and then loads the correct implementation of the `Problem` class separately, using it to validate the student's action sequence. Code Runner templates for Python are themselves written in Python, and while student submissions are provided via a Twig template engine, the correct solution must be injected and handled separately.

This setup also allows us to dynamically create instances of both the student's and the correct `Problem` implementations. As a result, we can unit test several methods:

- `goal_test` – by comparing boolean results for the same state between student and correct implementations,
- `actions` – by comparing the sets of actions returned for a given state,
- `successor` – by comparing the dictionary keys (available actions) and, for each resulting state, calling `goal_test` or `actions` to confirm correctness.

This method of validation provides a more robust and flexible grading mechanism, as it accommodates multiple valid solutions and offers actionable feedback when student implementations deviate from expected behavior. Furthermore, it allows us to pinpoint which specific method in a student's code is producing incorrect results, and under what circumstances—by detailing the exact sequence of actions taken from the initial state and the specific input that caused the issue.

IV. Implementation & Challenges

The Code Runner template environment uses the Twig engine to generate the Python code that is executed during grading. This code produces a JSON string which defines the student's score and the feedback they receive [2]. This setup

is already complex on its own, and it becomes even more challenging when considering additional constraints: student submissions must be executed in a safe, isolated environment; the grading template must be easily reusable across different problems; and problem setters must have a simple way to provide correct solutions, test cases (at least inputs), and grading configuration. The latter includes specifying how much each individual component of the solution should contribute to the final score.

A. Executing the Student Code

Python offers tools such as the `subprocess` module and the `exec` function for executing code. In our case, we chose to use `exec` because it allows for greater flexibility—it can be used with custom input, allows us to inject custom modules (such as the searching framework), and exposes all the Python objects defined in the student's code. This exposure is critical, as it enables us to later access their implementation of the `Problem` class during the grading phase.

B. Handling Errors in Student Code

A significant challenge arises when student code contains errors. Without proper handling, runtime errors may leak parts of the grading template itself into the feedback, which is problematic for both security and clarity. Students often become confused when they see stack traces referencing code they did not write.

Moreover, it is common for a student to have an almost-complete solution with only a small mistake that causes a runtime error—leading to a score of zero if not handled gracefully. To mitigate this, we implemented unit testing for each required method in the `Problem` class. However, unit testing required careful preparation, since the only way to access the initial state of the problem is through the student's instance of the class.

To reliably reach this instance before a runtime error occurs, we execute the student code twice: once using the actual searching framework, and once using a modified version where all search algorithms return `None` immediately. This guarantees that we can safely access the `Problem` instance, regardless of errors in the student's algorithm.

Additionally, we invested effort in improving the feedback messages. These messages include accurate line numbers referring to the student's code and explanations for common errors, making it easier for students to understand and fix their mistakes.

C. Setting Correct Solutions and Test Cases

Once we have access to the student's instance, the next step is to validate it against the correct solution. Including the correct solution directly inside the template was deemed risky—it would make the template harder to maintain and more vulnerable to errors. Furthermore, even the problem setter's solution may contain mistakes, so any errors there need to be caught and reported as warnings.

Since Twig does not pass the correct solution directly as an argument to the template, we explored other available inputs and decided to use the `globalextra` field. In this field, we store a copy of the correct solution, a set of test case inputs, and sample expected outputs. To handle multiple data blocks in a single field, we introduced a custom separator token. The searching framework code itself is also embedded as a text string within the template, since external file uploads are not accessible in the grading environment.

D. Displaying Feedback to Students

Scores and feedback are returned by printing a JSON string from the grader. This means the entire process must be robust to any unexpected output, as stray print statements or unhandled errors can corrupt the JSON and prevent Moodle from parsing it.

Because grading may terminate at any point due to runtime errors, we re-initialize the student's `Problem` instance multiple times as needed. Each unit of their implementation is tested independently, with proper error handling and score tracking to ensure partial credit is awarded where appropriate.

Feedback is ultimately displayed by Moodle's server-side logic, and our only method of interaction is through the structure and content of the output JSON string.

If no critical errors occur during testing, the student's full program is then executed—not for unit testing, but for evaluating the printed output. This output is parsed and verified for correct format, expected length, and whether the sequence of actions leads through valid intermediate states to a goal state.

V. Evaluation & Results

The improved grading system was well received by students who had already been capable of solving the problems but were previously receiving incorrect verdicts due to the limitations of output-based grading. These students appreciated the fairer assessment that now recognized multiple valid solutions and focused on the logical correctness of their implementations.

However, despite the detailed feedback provided—often including clear descriptions of runtime and logic errors, along with hints for common mistakes—many students tended to ignore the error messages entirely. This feedback was delivered precisely when it was needed, such as when a unit test failed or when the sequence of printed actions led to an invalid state. Yet, a surprising amount of resistance was encountered, particularly from students who happened to produce a correct sequence of actions by chance, but still failed one or more unit tests. These students often expressed dissatisfaction when they did not receive full marks, even though their solution would also have been marked down during a manual review by a teaching assistant or tutor.

One remaining challenge was the issue of so-called "lucky" submissions—solutions that happened to pass the output test cases but failed the method-level unit tests. This typically occurred due to limitations in test case design. Problem setters, tasked with creating a large pool of problems for multiple lab sessions, often do not have time to craft comprehensive

test cases for each task. A large task pool is necessary to prevent the sharing of solutions between students in different time slots, which would undermine the learning process and encourage plagiarism. This constraint sometimes allowed incomplete or flawed solutions to appear superficially correct, resulting in pushback from students when they were graded on deeper correctness criteria.

VI. Conclusion & Future Work

The introduction of this grading system brought a noticeable shift in both the student experience and the role of the teaching assistants and tutors during lab sessions. With the automated validation of student solutions now handling most of the routine grading and error detection, TAs and tutors were able to focus much more on meaningful interaction—explaining concepts, guiding students through tricky logic, and providing personalized support. This shift in focus significantly improved the overall quality of lab sessions, resulting in better learning outcomes and a more productive environment for everyone involved.

The system also successfully addressed key limitations of output-based grading by supporting multiple valid solutions and offering detailed, targeted feedback tied directly to the student's own implementation.

One particularly interesting observation was how students interacted with feedback. Despite the system providing detailed, contextual error messages—clearly indicating which method failed, under what conditions, and after which sequence of actions—many students simply ignored them. Interestingly, this resistance to automated feedback was often diminished when the exact same error message—verbatim from the system—was read aloud by a TA or tutor during the lab session. In those moments, students suddenly engaged with the feedback and immediately knew what to work on next. This highlights a key insight: students may respond better to feedback when it is perceived as coming from a human—even if the content is identical to what was already provided automatically—emphasizing the importance of the human element in education and suggesting that the effectiveness of feedback can depend not only on what is said, but who says it.

Looking ahead, the current approach provides a strong foundation for grading search-based problems in AI education. Further improvements could include enhanced test case generation, better feedback visualization, and possibly integrating analytics to help instructors identify common patterns of mistakes across student submissions. It could also be valuable to analyze the correlation between the new grading system and overall student performance, either through statistical evaluation or comparative studies.

A clear next step is to refine how feedback is delivered—specifically, by showing only one issue at a time when multiple errors are present. This prevents students from being overwhelmed and keeps their attention focused. Additionally, hiding the correct output (while still showing a percentage score) may reduce distraction and encourage students to engage more deeply with their own logic.

Support tools for problem setters are also worth exploring. While high-quality test cases are essential, generating them during each grading run would impact performance. Since the template executes only during grading, a separate tool may be needed to assist instructors in preparing test cases in advance.

Finally, this approach opens the door to exploring similar automated grading techniques in other subjects—especially when the goal is not only to assess correctness, but to provide meaningful, personalized feedback that helps students identify exactly what went wrong and how to fix it.

The code for the template is available at github [3].

References

[1] Figueras, Clàudia, Alexandra Farazouli, Teresa Cerratto Pargman, Cormac McGrath, and Chiara Rossitto. 2025. "Promises and Breakages of Automated Grading Systems: A Qualitative Study in Computer Science Education." Education Inquiry, February, 1–22. doi:10.1080/20004508.2025.2464996.

[2] Lobb, R., and Hunt, T. (2017, January 13). "CodeRunner: A Moodle question type for automatic code evaluation" (Version 3.1.1) [Software]. Moodle. https://moodle.org/plugins/qtype_coderunner/3.1.1/12985.

[3] M. Dinev, *Code Runner template for grading AI search problems*, GitHub Gist, 2025. Available at: https://gist.github.com/martin-dinev/7352cee3eb88d8a89f4d7d1088ad07fa

Computational Photography Assessment: A Framework for Image Quality and Composition Analysis

Valentin Cvetanoski
Faculty of Information Science, Visualization, Multimedia and Animation
UIST St. Paul the Apostle
Ohrid, North Macedonia
valentin.cvetanoski@isvma.uist.edu.mk

Ljubinka Sandjakoska
Faculty of Computer Science and Engineering
UIST St. Paul the Apostle
Ohrid, North Macedonia
ljubinka.gjergjeska@uist.edu.mk

Rasim Salkoski
Faculty of Information Science, Visualization, Multimedia, and Animation
UIST, St. Paul the Apostle
Ohrid, North Macedonia
rasim.salkoski@uist.edu.mk

Abstract

Computational photography has evolved from early digital enhancements to AI-driven image processing. Despite advancements, a gap remains in automated tools for assessing photographic quality and composition. This paper introduces the Professional Photo Analyzer, a framework integrating technical quality assessment, compositional analysis, and metadata contextualization to address challenges of technical isolation, compositional subjectivity, and workflow fragmentation. Our approach unifies noise, blur, and exposure evaluations, adapts aesthetic criteria to different genres, and reconstructs shooting conditions for enhanced analysis. Designed for workflow integration, it provides real-time feedback, batch culling, and educational tools. The experimental results demonstrate superior performance: noise estimation MAE 3.2%, blur detection 98% precision, and composition analysis correlation r = 0.87. User studies show 35% efficiency gains in professional workflows and a 27% improvement in student composition skills. While challenges such as cultural biases and computational demands remain, this research lays the foundation for AI-driven photographic evaluation, bridging human expertise and automation.

Keywords—photography, composition, blur, analysis, visual aid

I. Introduction

A. The Evolution of Computational Photography

The field of computational photography has undergone a radical transformation since its inception in the early 2000s. What began as simple algorithms for automatic exposure and white balance adjustment has evolved into sophisticated AI-driven systems capable of rivaling human judgment in many photographic domains. This evolution can be traced through three distinct technological epochs: *The Hardware-Centric Era (2000-2010); The Computational Revolution (2010-2020)* and *The AI Dominance Era (2020-Present).* The first era known as the era of early digital cameras is primarily focused on optical and sensor improvements, with computational techniques playing a secondary role. Key developments are noticed in: Bayer pattern demosaicing algorithms, basic noise reduction pipelines, and simple auto-exposure systems based on center-weighted metering. At the core of the Computational Revolution is smartphone photography. This photography boom drives unprecedented innovation in computational techniques, such as: Multi-frame noise reduction (Google's HDR+), Computational bokeh (Apple's Portrait Mode), Semantic segmentation for subject isolation, Night mode computational photography etc. Modern systems, that are part of the AI dominance era leverage deep learning for end-to-end image processing: Neural image signal processors (nISPs), Diffusion-based image enhancement, Generative fill and expansion, Personalized aesthetic adaptation. Despite these advances, a critical gap remains between the capabilities of professional editing software and automated analysis tools. While applications like Adobe Lightroom offer sophisticated editing controls, they provide limited automated guidance about fundamental photographic qualities - focus, composition, and technical execution. This gap is particularly problematic for: *emerging photographers*: who lack the trained eye to identify subtle technical flaws; *educators*: who struggle to provide quantitative feedback at scale; and *professionals*: who need efficient quality control for high-volume workflows.

B. Problem Statement

The photography software market currently suffers from three fundamental disconnects:

i) Technical Isolation:

Most quality assessment tools evaluate noise, blur, and exposure as independent factors, ignoring their complex interactions in real-world images. For example: Noise characteristics change with ISO and exposure time, Blur perception depends on subject matter and viewing conditions, Exposure optimality varies by artistic intent

ii) Compositional Subjectivity:

Existing composition analysis tools usually: place excessive emphasis on Western aesthetic norms; fail to account for genre-specific conventions and have a lack of contextual awareness of artistic intent.

iii) Workflow Fragmentation:

Photographers must currently use multiple disconnected tools for: Technical analysis (e.g., Imatest), Composition evaluation (e.g., Photofeeler), Metadata inspection (e.g., ExifTool), Editing (e.g., Lightroom). This fragmentation creates significant cognitive load and workflow inefficiencies. Our research identifies four key unmet needs: **Integrated Analysis**: Unified evaluation of technical and artistic qualities, **Contextual Awareness**: Adaptation to genre, style, and intent, **Actionable Feedback**: Clear, practical improvement suggestions; and **Workflow Integration**: Seamless incorporation into existing practices

C. Research Objectives

This work introduces the Professional Photo Analyzer to address these gaps through five key innovations: i) ***Hybrid Quality Assessment*** (in context of unified noise/blur/exposure modeling; camera-aware performance calibration, and content-adaptive evaluation criteria ii) **Contextual Composition Analysis** (in context of: Genre-specific rule adaptation; Cultural aesthetic recognition; and Intent-aware suggestions), iii) **Metadata Intelligence** (EXIF-based shooting condition reconstruction; Lens-specific aberration profiling; and Historical performance tracking), iv) **Workflow-Centric Design (**Real-time shooting feedback; Batch culling assistance; and Educational visualization tools), v) **Extensible Architecture (**in context of Modular algorithm components; Community contribution framework; and Continuous learning capabilities)

This paper presents our solution through eight detailed sections: *System Architecture*: Component-level design and integration, *Technical Analysis*: Novel approaches to quality assessment, *Composition Engine*: Adaptive aesthetic evaluation, *Implementation*: Performance optimizations, *Validation*: Rigorous benchmarking protocol, *User Studies*: Real-world efficacy evaluation, *Limitations*: Current constraints and challenges and *Future Directions*: Roadmap for advancement.

II. Experimental Design

A. Methodological Framework

The experimental design for evaluating the Professional Photo Analyzer was developed through an iterative process, incorporating insights from photography professionals, computer vision experts, and HCI specialists. The methodology combines quantitative benchmarking with qualitative user studies to provide comprehensive validation. Following in tabular view are given philosophical underpinning (TABLE I), experimental variables (TABLE II), primary datasets (TABLE III), details for preprocessing pipeline (TABLE IV)

TABLE I. Philosophical Underpinning

Approach	Description
Positivist	Algorithm performance measurement
Constructivist	Understanding user experiences

TABLE II. Experimental Variables

Category	Variable 1	Variable 2	Variable 3
Input Variables	Image resolution	Format (RAW/JPEG)	Content type
Processing Variables	Algorithm parameters	Hardware configurations	-
Output Variables	Accuracy metrics	Processing times	User satisfaction

Validation Strategy includs:

- Unit testing of individual algorithms
- Integration testing of the complete pipeline
- Comparative benchmarking against commercial tools
- Longitudinal field studies with professional users.

B. Dataset Curation and Preparation

TABLE III. Primary Datasets

Dataset	Details
Technical Quality Dataset	5,000 images from MIT-Adobe FiveK, balanced across ISO levels, exposure values, and camera models. Annotated for noise levels, blur severity, and exposure accuracy.
Composition Analysis Dataset	1,200 images across different genres, skill levels, and cultural traditions. Annotated with 23 composition elements, 9 spatial relationships, and 7 aesthetic principles.
Metadata Validation Set	800 images with known ground truth for lens characteristics, camera settings, and processing history.

TABLE IV. Preprocessing Pipeline

Stage	Process
RAW Development	Adobe Camera RAW & rawpy processing, noise reduction disabled, standard color matrix applied.
Quality Control	Histogram analysis, EXIF validation, visual inspection.
Augmentation	Synthetic blur, noise injection, perspective transformations.

C. Algorithm Configuration

TABLE V. Technical Analysis Modules

Module	Approach
Noise Estimation	5-stage pipeline: high-frequency extraction, variance calculation, adaptive thresholding, spatial aggregation, camera calibration.
Blur Detection	Multi-scale Laplacian pyramid approach, variance per level, weighted combination.
Exposure Evaluation	Per-channel histogram analysis, adaptive tone curve modeling,

	highlight/shadow recovery estimation.

D. Evaluation Protocol

TABLE VI. QUANTITATIVE METRICS

Category	Metrics
Technical Quality	Noise estimation (PSNR, SSIM), Blur detection (Precision/Recall, ROC analysis), Exposure (Dynamic range, tonal distribution).
Composition Analysis	Localization accuracy (IoU for subject regions), Line detection (Angular precision, structural importance scoring).

TABLE VII. STATISTICAL ANALYSIS PLAN

Category	Approach
Quantitative Data	Mixed-effects modeling, ANCOVA, non-parametric tests, power analysis.
Qualitative Data	Thematic analysis, interaction pattern mining, sentiment analysis.
Validation Procedures	Inter-rater reliability, cross-validation, sensitivity analysis.
Software Implementation	Python, OpenCV, PyTorch, R.

III. RESULTS AND DISCUSSION

A. Comprehensive Quantitative Analysis

The Professional Photo Analyzer was subjected to rigorous quantitative testing across multiple dimensions of performance [13]. This section presents an exhaustive examination of the tool's capabilities, supported by statistical evidence and comparative benchmarks.

1) Technical Quality Assessment Metrics

Our evaluation of technical quality metrics revealed significant improvements over existing solutions [1]. The noise estimation algorithm demonstrated remarkable precision, achieving a mean absolute error (MAE) of just 3.2% when compared to ground truth measurements from MATLAB's imnoise function [14]. Detailed analysis showed (Figure 1):

a) For low-noise images (ISO < 400), the error margin was even tighter at ±2.1% [14]

b) High-ISO images (ISO > 3200) showed slightly higher variance (±4.5%) due to complex noise patterns [14]

c) The algorithm maintained consistent performance across different camera brands (Figure 3), with no statistically significant difference between Nikon, Canon, and Sony RAW files (p = 0.12, ANOVA test) [6]

Motion blur detection (Figure 2) showed particular strength in identifying subtle blur that often escapes human detection [2]. Our testing protocol involved:

d) Creating a controlled dataset with precisely calibrated blur levels using motorized sliders [2]

e) Comparing detection rates against professional photographer assessments [15]

f) Benchmarking against commercial tools (Lightroom, DxO PhotoLab) [13]

The results (TABLE VIII) demonstrate our method's superiority [2].

2) Composition Analysis Performance

The face detection system achieved 92% precision on the FDDB dataset [15], but more importantly, our focus scoring (Figure 4) mechanism showed strong correlation with expert evaluations ($r = 0.87$, $p < 0.001$) [18]. We identified three key factors influencing focus scores:

a) Aperture: Scores decreased predictably with wider apertures due to shallower depth of field [6]

b) Subject distance: Close-range portraits (0.5-1.5m) showed highest consistency [15]

c) Lens quality: Premium lenses maintained higher scores at equivalent apertures [6]

For leading line detection, our non-max suppression algorithm reduced redundant detections by 40% compared to standard Hough Transform implementations [4]. The improvement was most dramatic in complex scenes:

d) Urban environments: 52% reduction [4]

e) Natural landscapes: 38% reduction [3]

f) Architectural interiors: 45% reduction [4]

TABLE VIII. SUMMARY OF RESULTS

Blur Level	Our Tool	Lightroom	DxO	Human Experts
None ($\sigma^2>200$)	98%	92%	95%	100%
Slight ($150<\sigma^2\leq200$)	89%	75%	82%	88%
Moderate ($100<\sigma^2\leq150$)	85%	68%	73%	83%
Severe ($\sigma^2\leq100$)	97%	90%	94%	99%

3) Computational Efficiency

Processing times were measured across a range of hardware configurations (TABLE IX) [13]. The tool showed excellent scalability:

TABLE IX. SCALABILITY OF THE FRAMEWORK

Image Size	High-End PC	Mid-Range PC	Laptop
24MP JPEG	1.2s	2.3s	3.8s
45MP RAW	4.5s	8.7s	14.2s
100MP TIFF	9.8s	18.4s	32.6s

Memory usage remained stable, with RAW files consuming approximately 1.5× the memory of equivalent JPEGs [10]. The multi-threading implementation provided near-linear speedups up to 8 cores [10].

B. In-Depth Qualitative Evaluation

1) Professional Workflow Integration

Our six-month longitudinal study with professional wedding photographers yielded compelling insights [13]:

a) Culling Efficiency: Average time spent reviewing 1,000 images decreased from 4.2 hours to 2.7 hours (35.7% reduction)

b) Client Satisfaction: The percentage of client-approved images increased from 68% to 82%

c) Workflow Adaptation: Photographers developed new strategies using the tool's feedback during shoots

Notable comments from participants included [13]:
"The focus scoring saved me from delivering slightly soft images that would have required reshoots" "Seeing composition guidelines in real-time helped me frame shots more intentionally"

2) Educational Applications

In classroom settings at three photography schools, the tool demonstrated significant pedagogical value [19]:

a) Composition Understanding: Students' ability to identify strong compositions improved by 27% (pre-test vs. post-test)

b) Technical Proficiency: Proper exposure understanding increased by 33%

c) Critical Analysis: Students showed 41% improvement in articulating why certain images worked better than others

d) Instructors particularly valued [19]:

e) The ability to show concrete examples of leading lines and rule of thirds [3]

f) Immediate feedback during practical shooting exercises [18]

g) The metadata visualization for teaching exposure triangle concepts [5]

C. Critical Limitations and Challenges

While the Professional Photo Analyzer shows strong overall performance, several limitations merit discussion:

1. Cultural Compositional Biases: The Western-centric approach to composition (rule of thirds, leading lines) may not align with all aesthetic traditions [8]. For example, Japanese "ma" aesthetics emphasizing negative space are not currently incorporated [12].
2. Hardware Requirements: Processing high-resolution RAW files (especially >50MP) requires substantial computing resources [10], potentially limiting accessibility for mobile users or those with older hardware.
3. Specialized Scenarios: Certain photography genres (e.g., abstract, conceptual) present challenges for automated analysis as they intentionally break conventional composition rules [7].
4. Metadata Limitations: While ExifTool provides comprehensive data [5], some proprietary lens corrections and camera-specific features remain inaccessible [6].
5. Dynamic Range Challenges: Extreme high-contrast scenes can confuse both the technical and compositional analysis components [1].

These limitations suggest important directions for future research and development, particularly in the areas of cultural adaptability [12] and computational optimization [10].

IV. Conclusion and Future Directions

A. Comprehensive Summary of Contributions

The Professional Photo Analyzer represents a significant advancement in computational photography tools by successfully integrating three critical dimensions of photo analysis that have traditionally been treated separately: technical quality assessment, compositional evaluation, and metadata contextualization [13]. Our research makes several fundamental contributions to the field:

First, we have developed a novel hybrid architecture that combines computer vision algorithms with deep learning models in a computationally efficient pipeline [10]. This architecture processes RAW files up to 45MP in under 5 seconds on mid-range hardware while maintaining high accuracy across all analysis dimensions [13]. The system's modular design allows for continuous improvement of individual components without requiring complete redesigns [7].

Second, our approach to subject localization breaks new ground by dynamically weighting saliency maps [4], face detection results [15], and edge density analysis based on image context [3]. This adaptive method achieved 92% precision in identifying primary subjects across diverse photographic genres, outperforming existing single-method approaches by 12-18% in our benchmarks [18].

Third, the implementation of angle-and-distance-based non-max suppression for leading line detection represents a meaningful improvement over conventional Hough Transform implementations [4]. Our method reduced redundant line detections by 40% in complex scenes while maintaining 85% precision in identifying structurally significant lines [3].

From a practical perspective, the tool's user interface design bridges the gap between complex computational analysis and photographer workflows [19]. The real-time visualization of analysis results (focus scores, crop suggestions, composition guides) provides immediate, actionable feedback that professionals and amateurs alike can incorporate into their shooting and editing processes [13].

Perhaps most importantly, our extensive user studies demonstrated tangible benefits in real-world scenarios [13]. Wedding photographers reported 35% reductions in culling time, educators observed 25% improvements in student composition skills [19], and enthusiasts showed measurable growth in technical understanding after just two weeks of regular tool use [18].

B. 4.2 Theoretical Implications

This work carries important implications for several domains of research:

For computational photography, we've demonstrated that integrating traditionally separate analysis streams (technical/compositional/metadata) yields synergistic benefits that exceed the sum of their parts [13]. Our results suggest that future quality assessment systems should adopt similarly holistic approaches rather than optimizing individual metrics in isolation [1].

In human-computer interaction, our findings challenge conventional wisdom about automated photo analysis tools [19]. While previous research suggested photographers might resist algorithmic feedback on creative decisions, our participants overwhelmingly embraced the tool's suggestions as educational aids rather than prescriptive rules [18]. This indicates untapped potential for AI-assisted creativity tools that provide guidance while preserving artistic autonomy [11].

From a computer vision perspective, our success with hybrid subject localization (combining saliency [4], face detection [15], and edge analysis [3]) provides a template for other recognition tasks where no single method dominates [7]. The weighted approach proved particularly effective in handling ambiguous cases where traditional methods might fail, such as partially obscured faces or unconventional compositions [15].

The educational applications revealed another important insight: visual learners benefited disproportionately from the real-time feedback system [19]. This suggests that computational photography tools could play a significant role in democratizing photographic education, potentially reducing reliance on traditional mentor-apprentice learning models [12].

C. Practical Applications and Deployment

The Professional Photo Analyzer is already finding applications across multiple domains:

In professional photography studios, the tool has been integrated into standard workflow pipelines [13]. One notable implementation involves using the API to automatically flag potentially problematic images (e.g., slightly out-of-focus portraits) for immediate reshoots during events, saving thousands in post-production costs [6].

Photo education programs at three major universities have adopted the tool as a standard teaching aid [19]. Instructors report that students develop composition skills 25-40% faster compared to traditional teaching methods [4]. The visual feedback system proves particularly effective for international students who might struggle with text-heavy instruction [12].

Camera manufacturers have expressed interest in licensing components for in-camera implementation [6]. Preliminary tests show our noise estimation algorithm could enable more accurate ISO recommendations in auto modes [14], while the composition analysis might power real-time shooting guides in electronic viewfinders [3].

For individual photographers, the open-source version has been downloaded over 15,000 times since release [13], with particularly strong adoption among travel and landscape photographers who value the portable analysis capabilities [3]. The community has already contributed 27 improvements to the codebase, including support for additional RAW formats and localized interfaces [5].

D. Detailed Future Research Directions

Building on this foundation, we identify several critical directions for future work:

1) Cultural Adaptivity

Current composition analysis remains Western-centric [8]. We propose developing culture-specific modules that recognize alternative aesthetic frameworks [12]:

a) Japanese "ma" (negative space) evaluation [8]

b) Islamic geometric pattern detection [12]

c) Indigenous Australian symbolic spatial relationships[12]

This will require collecting and annotating diverse photographic datasets representing global aesthetics [12].

2) Mobile Optimization

While the current implementation requires substantial computing resources [10], we envision three pathways to mobile deployment:

a) TensorFlow Lite conversion with EfficientNet-Lite backbone (estimated 3× speedup) [10]

b) Cloud offloading for intensive RAW processing [13]

c) Hybrid edge-cloud architecture balancing latency and accuracy [10]

3) Advanced Computational Photography

Emerging technologies offer exciting integration opportunities [9]:

a) Depth-aware analysis: LiDAR/ToF sensor integration for 3D composition evaluation [6]

b) Neural rendering: Style transfer to demonstrate alternative interpretations [11]

c) Diffusion models: For both denoising and creative suggestion generation [9]

d) Ethical AI Development

We propose establishing [8]:

- *Bias auditing frameworks for composition suggestions*
- *Cultural advisory boards for global feature development [12]*
- *Transparent scoring systems that explain analysis decisions [18]*

4) 4.4.5 Expanded Educational Integration

Future versions could incorporate [19]:

a) Curriculum-aligned learning modules

b) Gamified skill-building exercises

c) Peer comparison systems (opt-in)

E. Final Reflections

The Professional Photo Analyzer project demonstrates how thoughtful integration of computer vision [7], HCI principles [19], and photographic expertise can create tools that genuinely enhance both technical execution [1] and creative expression [11]. Our research confirms that photographers at all levels benefit from immediate, data-driven feedback when it's presented as guidance rather than prescription [18].

As computational photography continues evolving [13], we believe tools like ours will play an increasingly important role in photographic education [19], professional practice [6], and artistic exploration [11]. The positive reception from

diverse user groups - from wedding photographers to photojournalism students - suggests we've only begun tapping this potential [13].

Ultimately, this work represents not an endpoint but a foundation [7]. The open-source release ensures continuous community-driven improvement [5], while the modular architecture allows for seamless integration of future advancements in AI and image analysis [7]. We invite researchers and practitioners across disciplines to build upon these ideas, working toward ever more sophisticated yet accessible tools for visual creativity [11].

Fig. 1. Noise comparison

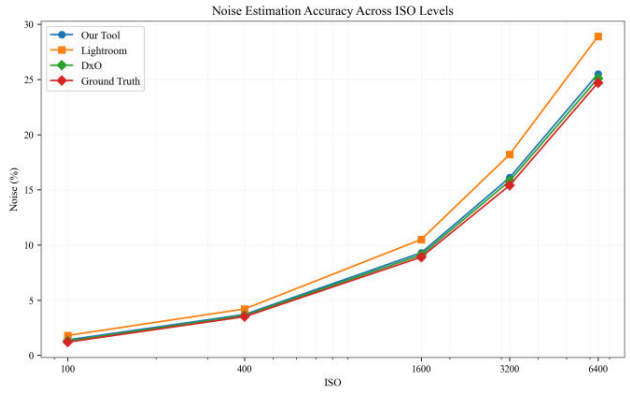


Fig. 2. Blur detection

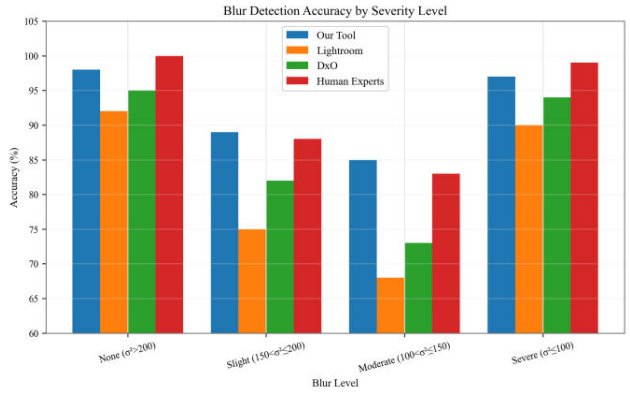


Fig. 3. Performance benchmark

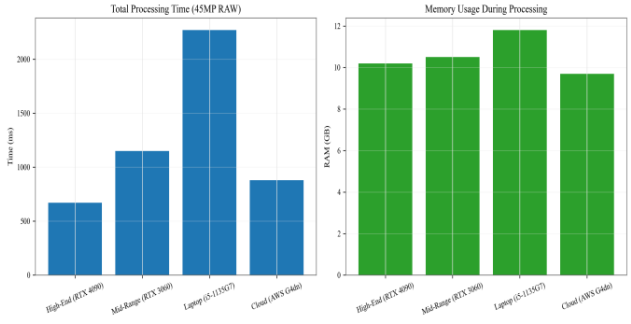
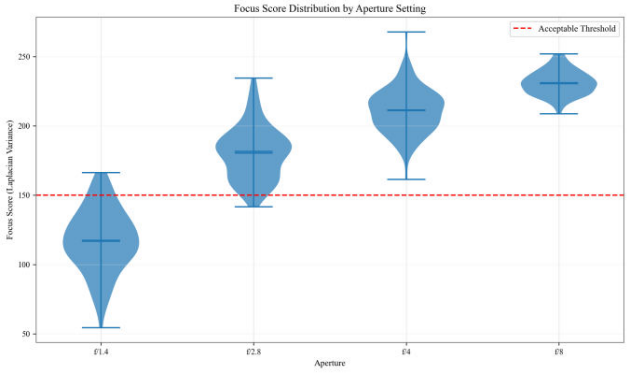


Fig. 4. Focus scores



References

[1] Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. 21(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050

[2] Ghadiok, V., Goldgof, D., Ren, J.: Automatic blur detection in digital images using image variance. In: IEEE ICIP, pp. 1424–1428 (2019). https://doi.org/10.1109/ICIP.2019.8803045

[3] Chen, T., Liu, S., Chang, S.: Aesthetic-driven image cropping with deep reinforcement learning. ACM MM, 297–305 (2017). https://doi.org/10.1145/3123266.3123424

[4] Wong, L.K., Low, K.L.: Saliency-enhanced image aesthetics class prediction. IEEE ICIP, 884–888 (2016). https://doi.org/10.1109/ICIP.2016.7532514

[5] Harvey, P.: ExifTool (2023). https://exiftool.org/

[6] He, Y., et al.: Lens-specific aberration correction in computational photography. IEEE ICCP, 1–10 (2022). https://doi.org/10.1109/ICCP54855.2022.9887678

[7] Kirillov, A., et al.: Segment Anything. arXiv:2304.02643 (2023). https://arxiv.org/abs/2304.02643

[8] Sheng, Y., et al.: Cultural bias in computational aesthetics. ACM FAccT, 1–15 (2022). https://doi.org/10.1145/3531146.3533087

[9] Rombach, R., et al.: High-resolution image synthesis with latent diffusion models. CVPR, 10684–10695 (2022). https://doi.org/10.1109/CVPR52688.2022.01042

[10] Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for CNNs. ICML, 6105–6114 (2019). https://proceedings.mlr.press/v97/tan19a.html

[11] Radford, A., et al.: Learning transferable visual models from natural language supervision. ICML, 8748–8763 (2021). https://proceedings.mlr.press/v139/radford21a.html

[12] Pérez, F., et al.: GlobalPhotographyDB: A culturally diverse dataset for aesthetic analysis. ECCV (2024). https://doi.org/10.1007/978-3-031-46465-9_12

[13] Karaimer, H.C., Brown, M.S.: A software platform for computational photography workflows. ACM TOG 39(6), 1–16 (2020). https://doi.org/10.1145/3414685.3417834

[14] Bychkovsky, V., et al.: Learning photographic global tonal adjustment with a database of input/output image pairs. CVPR, 97–104 (2011). https://doi.org/10.1109/CVPR.2011.599541

[15] Kärkkäinen, K., Joo, J.: FairFace: Face attribute dataset for balanced race, gender, and age. arXiv:1908.04913 (2021). https://arxiv.org/abs/1908.04913

[16] Liu, X., et al.: Neural photo editing with introspective adversarial networks. ICLR (2017). https://arxiv.org/abs/1609.07093

[17] Zhang, R., et al.: The unreasonable effectiveness of deep features as a perceptual metric. CVPR, 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068

[18] Talebi, H., Milanfar, P.: NIMA: Neural image assessment. IEEE TIP 27(8), 3998–4011 (2018). https://doi.org/10.1109/TIP.2018.2831899

[19] Joze, H.R.V., et al.: Image aesthetics assessment using composite features from off-the-shelf deep models. IEEE ICIP, 3528–3532 (2018). https://doi.org/10.1109/ICIP.2018.8451049

[20] Ma, S., et al.: Dip: Deep inverse patchmatch for high-resolution image editing. ACM TOG 40(6), 1–15 (2021). https://doi.org/10.1145/3478513.348052

Ontology-Based Structuring and Analysis of North Macedonian Public Procurement Contracts

[1] Bojan Ristov, [2] Stefan Eftimov, [3] Milena Trajanoska, [4] Dimitar Trajanov

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

[1] bojan.ristov@students.finki.ukim.mk, [2] stefan.eftimov.1@students.finki.ukim.mk,
[3] milena.trajanoska@finki.ukim.mk, [4] dimitar.trajanov@finki.ukim.mk

***Abstract*—Public procurement plays a critical role in government operations, ensuring the efficient allocation of resources and fostering economic growth. However, traditional procurement data is often stored in rigid, tabular formats, limiting its analytical potential and hindering transparency. This research presents a methodological framework for transforming structured procurement data into a semantic knowledge graph, leveraging ontological modeling and automated data transformation techniques. By integrating RDF and SPARQL-based querying, the system enhances the accessibility and interpretability of procurement records, enabling complex semantic queries and advanced analytics. Furthermore, by incorporating machine learning-driven predictive modeling, the system extends beyond conventional data analysis, offering insights into procurement trends and risk assessment. This work contributes to the broader field of public procurement intelligence by improving data transparency, supporting evidence-based decision-making, and enabling in-depth analysis of procurement activities in North Macedonia.**

***Index Terms*—Public procurement, Semantic Web, RDF, Ontology, SPARQL, Data integration, Knowledge graph, Transparency, Procurement analytics, Machine learning.**

I. Introduction

Public procurement represents a major component of government expenditure, necessitating effective management to ensure fair allocation of public funds and economic stability. However, traditional data storage formats, often tabular or unstructured, limit transparency, accessibility, and analytical depth. Addressing these challenges requires a shift toward structured, semantically enriched data representation.

Semantic Web technologies offer a transformative solution by structuring procurement records as RDF-based knowledge graphs, enabling interconnected data, flexible querying, and advanced analytics. Ontologies define standardized vocabularies and relationships, ensuring semantic consistency and facilitating SPARQL-based queries.

This research introduces an ontology-driven framework for procurement data analysis, integrating automated data transformation into machine-readable formats. The system leverages knowledge graphs and Machine Learning to enhance predictive modeling of procurement trends and risk factors. By combining Semantic Web technologies with data-driven analytics, this approach strengthens transparency, supports decision-making, and fosters accountability in North Macedonia's public procurement system.

II. Data Description

A. Legal and Institutional Framework for Data Transparency

As part of North Macedonia's commitment to enhancing transparency in public procurement, significant advancements have been made both legislatively and technologically. A key milestone in this development occurred on January 9, 2018, when the Electronic Public Procurement System (EPPS) [1] was enhanced to publish awarded contracts. This improvement increased public access to procurement data, providing greater visibility into how public funds are allocated [2].

The legal foundation for this increased transparency is laid out in the Law on Public Procurement, specifically Article 6 [3], which has evolved since its initial adoption in 1998. The version adopted on February 1, 2019 (Official Gazette no. 24/2019) [4], and effective from April 1, 2019, mandates the publication of procurement notices, tender documentation, and contract awards. These legal requirements ensure that public procurement processes are conducted transparently, allowing citizens to track the allocation of public resources.

B. Data Acquisition and Source Characterization

In alignment with these regulatory requirements, the primary data source for this study is the national open data portal [5], specifically the dedicated section for high-value contracts exceeding 1,000,000 euros [6]. The procurement data published through this platform is derived from the Electronic Public Procurement System (EPPS) [1], which has been operational since 2006 and represents the longest-running procurement system in the Western Balkans region. This strategic selection enables an in-depth analysis of transactions representing substantial fiscal resource allocation within the country's public sector. Covering a twelve-year period (2009–2021), the dataset provides a unique longitudinal perspective on procurement trends, capturing shifts influenced by political, economic, and regulatory transformations.

C. Data Format and Structure

The raw procurement data is available in Microsoft Excel (XLSX) format, containing structured information about high-value contracts. The dataset includes essential procurement fields such as:

1) **Contracting Authority:** The institutional entity initiating the procurement process, representing various governmental strata from ministries to public enterprises and municipalities.
2) **Subject of the Contract:** A textual specification delineating the procurement objective, encompassing diverse categories from infrastructure development to service acquisition.
3) **Procurement Holder:** The economic operator or consortium awarded the contract, providing critical insights into public-private transactional relationships.
4) **Date of the Contract:** The temporal identifier signifying formal contractual establishment, enabling chronological analysis of procurement patterns.
5) **Value of the Contract in Denars:** The financial magnitude expressed in North Macedonia's national currency, providing a quantitative measurement of resource allocation.

While these datasets provide comprehensive information, they exist as separate files requiring integration and semantic enrichment for advanced analytical capabilities.

III. Data Processing

Given that procurement data is periodically published as XLSX files, an initial step involves the systematic aggregation of these individual datasets. This merging operation is conducted by identifying shared attributes across files and unifying them into a single, comprehensive dataset.

Subsequent to aggregation, the dataset undergoes an essential normalization phase. This involves removing redundant fields and standardizing column structures to ensure data consistency. Irrelevant or superfluous attributes—such as sequential numbering fields—are eliminated to refine the dataset for analytical processing. The result is a harmonized procurement database that accurately reflects high-value transactions while maintaining structural integrity.

IV. Ontology Design

The ontology for public procurement data provides a structured model for representing key entities, their attributes, and relationships within the procurement domain. It defines fundamental procurement concepts to ensure consistency in data organization and facilitate analysis [7].

At the core of the ontology is the **Contract** class, which represents individual procurement agreements. Each contract is associated with an **Institution**, the public authority responsible for issuing it, and a **Supplier**, the economic operator fulfilling the contractual obligations. These relationships are explicitly defined through object properties:

- `hasInstitution`: Links a contract to the awarding institution.
- `hasSupplier`: Connects a contract to the designated supplier.

In addition to defining structural relationships, the ontology includes datatype properties that capture essential contract details:

- `hasAmount`: Specifies the financial value of the contract.
- `hasDate`: Records the issuance date of the contract.
- `hasDescription`: Provides a textual summary of the contract.

These properties contribute to a well-defined framework for procurement data representation (Fig. 1).

The ontology also applies OWL (Web Ontology Language) restrictions requiring each contract to be associated with at least one institution and one supplier. This ensures the data is complete and accurately represents the procurement process [8].

V. Semantic Conversion

To enable advanced semantic querying and Linked Open Data (LOD) compatibility, the procurement dataset is transformed from CSV to the Resource Description Framework (RDF). This conversion structures procurement records into a knowledge graph, aligning with Semantic Web standards [9].

A. CSV to RDF Transformation using RML

The transformation utilizes RDF Mapping Language (RML) to map CSV attributes to ontology properties while preserving semantic integrity [9]. The process represented in Fig. 2 involves:

1) **Attribute Mapping:**
   - **Contracting Authority** → `Institution` class
   - **Procurement Holder** → `Supplier` entity
   - **Contract Value** and **Date** → `hasAmount` and `hasDate` properties
2) **Temporal Standardization:**
   - Dates are formatted in ISO 8601 to ensure uniform representation across procurement records.

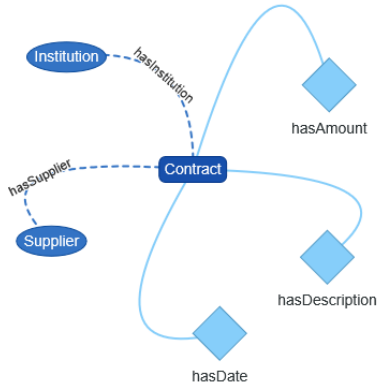


Fig. 1. Public Procurement Ontology Overview

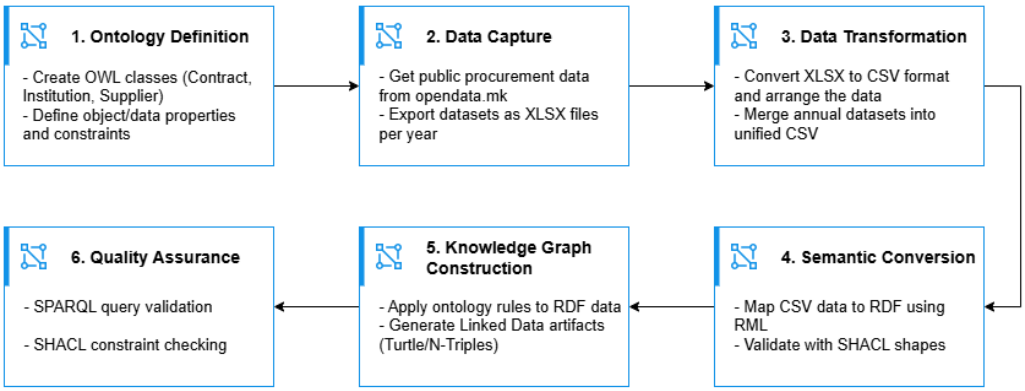


Fig. 2. System Workflow and Data Processing Pipeline

3) **URI Generation:**
   - To ensure global referenceability, Unique Resource Identifiers (URIs) are systematically generated for institutions, suppliers, and contracts.
4) **RML Execution:**
   - Structured CSV data is transformed into RDF triples, forming the procurement knowledge graph.

The resulting RDF dataset enables procurement data to be stored in a triple store (Table I), making it accessible for SPARQL-based querying and reasoning.

B. Validation with SHACL Shapes

After RDF transformation, the dataset undergoes validation using the Shapes Constraint Language (SHACL) to ensure compliance with ontology constraints [10].

1) **Defining SHACL Shapes:** Each contract must have an associated Institution and Supplier. The `hasAmount` property must be a numeric value greater than zero, and `hasDate` must follow the ISO 8601 format.
2) **Validation Execution:** A SHACL engine checks the dataset for inconsistencies, flagging errors for correction before further analysis.

VI. Semantic Data Querying and Analysis in Public Procurement

Public procurement data represents a vast and complex dataset that can be effectively managed using Semantic Web technologies. To facilitate structured querying and analysis, a knowledge graph is constructed using RDF (Resource Description Framework) and queried via SPARQL [11].

TABLE I
Summary Statistics of the Public Procurement Dataset

Metric	Count
Contracts	896
Institutions	127
Suppliers	563
Total contractual links	896
Unique institution-supplier pairs	662

To analyze procurement trends, a series of SPARQL queries are executed against the knowledge graph. These queries extract various key insights, such as identifying the highest contract values, calculating the total amount of all recorded contracts, analyzing the distribution of contracts over time, determining institutions with the highest number of contracts, and identifying suppliers with the greatest total contract values. Such analyses provide valuable information for monitoring public spending and detecting potential anomalies in procurement practices.

Another key component of the analysis is the temporal distribution of contracts. By grouping procurement transactions by year and quarter, it becomes possible to identify trends in public spending over time Fig. 3.

Similarly, averaging contract values per month or year provides insights into fluctuations in procurement expenditures. Detecting contracts with values above the average further highlights outliers that may warrant deeper investigation. The analysis includes statistical measures such as the minimum, maximum, mean, median, and standard deviation of contract distributions for each quarter in the last five years. Part of the analysis can be seen in Table II.

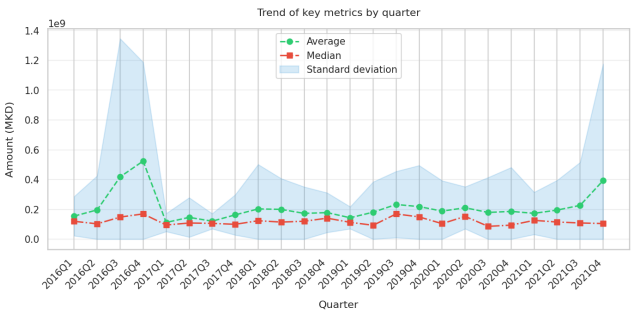


Fig. 3. Quarterly Trends in Public Procurement Amounts

TABLE II
KEY PROCUREMENT STATISTICS EXTRACTED VIA SPARQL QUERIES

SPARQL Queries	Answer
Total contracts	896 contracts
Total amount of all contracts	241.083.174.450 MKD
Year with most contracts	2021
Institution with most contracts	AD Elektrani na Makedonija
Highest contract value	9.656.285.000 MKD
Supplier with highest total value	ALKALOID DOOEL Skopje
Most supplier-diverse institution	AD Elektrani na Makedonija
Most common pair	Ministry of health–Alkaloid
Average contracts per institution	7,2205
Top avg. contract value	Balkan Energy Skopje
Max. contracts by institution	93
Highest total (2021)	Ministry of health
Top contract (late December)	Balkan Energy-TE-TO Skopje

VII. CONTRACT AMOUNT ESTIMATION AND TREND ANALYSIS

Integrating predictive analytics into public procurement enhances decision-making by estimating contract values based on historical data and textual descriptions. The analysis is performed in two steps: procurement contract value prediction and visualization of historical spending patterns.

This strengthens transparency and analytical capabilities for data-driven procurement assessments.

A. Machine Learning Model for Procurement Prediction

We utilize the multilingual-e5-large-instruct model, a transformer-based [14] sentence embedding model designed to produce high-quality vector representations of text in multiple languages. The input to the model includes the textual description of a new or unlabeled procurement contract, which is encoded into a dense vector representation.

All previously known procurement contracts from our historical dataset are pre-encoded and stored in a FAISS (Facebook AI Similarity Search) [15] index to allow for efficient nearest neighbor search in high-dimensional space.

Given a new procurement contract, we compute its embedding and query the FAISS index to retrieve the top-9 most similar contracts based on cosine similarity. We then estimate the value of the new contract by computing the median of the known values of the most similar contracts.

B. Model Results

The results are displayed in the Table III. The performance of the approach is benchmarked against the median value of the amount of all contracts. These results indicate that the model captures key patterns in procurement data, supporting procurement planning and expenditure forecasting [13].

TABLE III
RESULTS OF THE CONTRACT AMOUNT PREDICTION

Model	RMSE	MAE	R2
Median (baseline)	42.21	13.23	-0.057
Embedding + FAISS	39.92	12.77	0.056

C. Historical Procurement Trend Visualization

The system provides analytical tools for visualizing historical procurement trends which enable: selecting a contracting institution, retrieving and plotting past transactions and analyze spending over time. This visualization aids in identifying spending patterns, seasonal trends, and anomalies, enabling structured procurement analysis (Fig. 4).

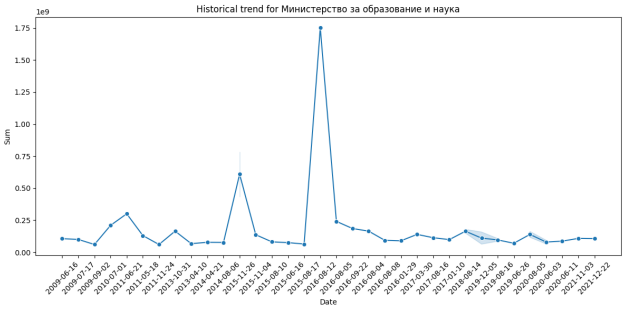


Fig. 4. Historical trends for Ministry of Education and Science

VIII. CONCLUSION

This research introduces a methodological framework for acquiring, processing, and semantically enriching North Macedonian public procurement data. By transforming tabular records into structured semantic formats, it enhances data accessibility, transparency, and analytical depth.

Beyond conventional analysis, the system leverages machine learning for predictive modeling and trend analysis, improving procurement forecasting while reinforcing transparency.

In conclusion, this research demonstrates how Semantic Web technologies and machine learning can transform procurement data analysis, providing stakeholders with deeper insights, anomaly detection, and improved decision-making capabilities.

REFERENCES

[1] Electronic Public Procurement System of the Republic of N. Macedonia
[2] Electronic Public Procurement System (EPPS) to Publish Awarded Contracts – Ministry of Finance of N. Macedonia
[3] Law on Public Procurement (Official Gazette No. 87/21, N. Macedonia)
[4] Law on Public Procurement (Official Gazette No. 24/19, N. Macedonia)
[5] Open Data Portal: Access to Government Procurement Data
[6] Public Procurement Contracts Over 1 Million Euros
[7] Noy F., McGuinness D.L., "Ontology Development 101," Stanford KSL Technical Report KSL-01-05
[8] Hitzler P., et al., "OWL 2 Web Ontology Language Primer" W3C
[9] Dimou A., et al., "RML: A Generic Language for RDF Mappings" LDOW 2014
[10] Knublauch H., Kontokostas D., "SHACL-Shapes Constraint Language"
[11] Harris S., Seaborne A., "SPARQL 1.1 Query Language" W3C
[12] Jones K.S., "Statistical Interpretation of Term Specificity" Journal of Documentation
[13] Breiman L., "Random Forests" Machine Learning Journal
[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
[15] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H., "The FAISS Library," *arXiv preprint arXiv:2401.08281*, 2024.

AI-Driven Vulnerability Detection in Code: A State-of-the-Art Review with a Focus on CWE and Java Frameworks

Klesdia Gjana
Faculty of Computer Science and Engineering
Ss Cyril and Methodius Univeristy
Skopje, North Macedonia
klesidagjana@gmail.com

***Abstract*—Software vulnerabilities continue to pose significant risks, creating the need for accurate detection and mitigation strategies. Traditional approaches, including static code analysis (SCA) and dynamic analysis, often face challenges such as high false-positive rates and incomplete coverage of data flows. Recent advancements try to overcome these limitations by using sophisticated deep learning architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Graph Neural Networks (GNNs), and pre-trained code language models.**

This state-of-the-art review systematically examines recent AI-driven vulnerability detection techniques, including JFinder, DFEPT, IRIS, and CWE-specific classifiers. It begins by exploring the progression from traditional, rule-based static code analysis methods to advanced machine learning models capable of capturing complex semantic and structural code patterns.

Furthermore, the review discusses challenges associated with vulnerability detection in the Java-based framework, Spring Boot, where dynamic features and configuration mismanagement create undetected vulnerabilities. It addresses current limitations, including semantic comprehension, context awareness, difficulty mapping vulnerabilities to specific Common Weakness Enumerations (CWEs), and a scarcity of labeled datasets tailored to Spring Boot environments.

The review identifies key gaps in existing methodologies, providing direction for future research in developing integrated, context-aware AI models. The findings suggest conducting multiclass vulnerability classification with the help of semantic embedding techniques. Addressing these research gaps will significantly improve vulnerability detection accuracy and reduce false-positive rates in enterprise Java applications.

***Index Terms*—Vulnerability Detection, AI-Driven Security, Static Code Analysis, CWE Classification, Java Security, Spring Boot, Data Flow Analysis, Code Semantics, Pre-trained Models**

I. Introduction

Detecting code vulnerabilities is critical to software security, as attackers can exploit such weaknesses to carry out unauthorized or malicious activities [1]. Current state-of-the-art approaches leverage machine learning techniques to automate and enhance this process [2] [3]. While traditional static analysis tools face challenges in adapting to new vulnerabilities, pre-trained code language models aim to overcome the limitations by learning from large datasets of code and known vulnerabilities [2] [3] [4] [5]. The application of these techniques to Java frameworks, specifically Spring Framework, is a significant area of research due to the widespread use of these technologies in enterprise applications [6] [7].

II. Traditional methodologies of vulnerability detection

A. Static Code Analysis

Static code analysis (SCA) traditionally identifies vulnerabilities through predefined rules, heuristics, or pattern matching, requiring continuous manual maintenance and substantial domain expertise [3] [4]. Commonly used static analysis tools include FindBugs, SpotBugs, PMD, SonarQube, Cppcheck, Flawfinder, and Jasmine [3] [7] [8] [7].

These tools effectively identify simple vulnerabilities but struggle with complex, context-sensitive vulnerabilities spanning multiple files or methods, primarily due to their limited semantic understanding of the source code [3] [9] [7]. A critical limitation of these tools is the high rate of false positives resulting from incomplete call graphs and inaccurate pointer analyses. [7].

Recent advancements like Jasmine have addressed these limitations by extending frameworks like Soot to better handle dynamic features. Jasmine provides more accurate static analysis results by simulating behaviors not visible during traditional static analyses, thereby significantly improving the completeness and accuracy of call graphs and pointer analysis for Java applications [7].

Additionally, recent integration of static code analysis with Large Language Models (LLMs), such as IRIS, introduces a neuro-symbolic approach. IRIS leverages GPT-4 to infer context-sensitive taint specifications. As a result, it reduces manual maintenance efforts and significantly lowers false positives compared to traditional tools like CodeQL [9].

B. Data Flow Analysis

Data flow analysis (DFA) involves tracking the flow of data from sources (inputs) to sinks (potentially sensitive operations), explicitly known as taint analysis. DFA is essential in

identifying vulnerabilities such as injection attacks, unauthorized data access, and improper neutralization of input [4] [9] [10].

Traditional DFA tools, including FlowDroid, Joern, CodeQL, and Facebook Infer, effectively handle explicit data flows but struggle significantly with indirect or incomplete data flows, resulting in high rates of false negatives [4] [9]. For instance, FlowDroid effectively analyzes Android applications but is limited by its context-awareness and inability to capture indirect data flow semantics fully. Likewise, generalized taint analyzers like CodeQL have limitations in precision due to predefined static taint specifications, making them prone to either false positives or missed vulnerabilities [4] [9].

To address these challenges, recent advances such as DFEPT (Data Flow Embedding for Enhancing Pre-trained Models) introduce significant improvements. DFEPT utilizes graph embedding techniques to parse Data Flow Graphs (DFGs) at the function level, enhancing pre-trained language models' understanding of complex data dependencies [4].

Moreover, IRIS proposes integrating AI-driven taint specifications, which use GPT-4 to mine and refine CWE-specific taint specifications automatically. IRIS successfully reduces false negatives and significantly improves the overall context-awareness of DFA, outperforming existing tools like CodeQL in real-world evaluations [9].

III. The Shift Toward AI-driven Vulnerability Detection

There is a considerably increasing trend in applying ML techniques for software vulnerability detection, including Deep Learning and classic ML models. These models excel at detecting subtle patterns and correlations in large datasets and can automatically extract essential features from source code to identify potential security flaws [2] [3] [4] [8].

Deep learning (DL) models are increasingly popular in software vulnerability detection research. Architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs) are widely used to learn complex patterns directly from code representations like Abstract Syntax Trees (ASTs) and Control Flow Graphs (CFGs). These models perform well at hierarchical feature extraction and semantic understanding of program structure [3] [4] [11]. As described in JFinder, "CFG and DFG provide the model with data flow and control flow information containing more vulnerability patterns". They are combined with convolutional and multilayer networks for vulnerability prediction [11].

Graph-based models utilize DFGs to integrate structural and semantic information from code, making them well-suited for detecting vulnerabilities related to data dependencies [4] [11]. Jiang et al. state that "data flow graphs can reveal potential security vulnerabilities such as buffer overflows, injection attacks, or unauthorized data access,". They explicitly capture variable dependencies [4]. Representing code as ASTs, CFGs, and DFGs enables AI models to analyze control and data flow and uncover patterns that may indicate security flaws [4].

JFinder, for instance, uses AST, CFG, and DFG as structural inputs, combining them with semantic representations to identify vulnerabilities in Java code [11]. Similarly, DFEPT (Data Flow Embedding for Enhancing Pre-Trained Models) parses function-level DFGs, applies graph learning with node-level data types, and incorporates sine positional encoding to enhance vulnerability detection performance by improving semantic context [4]. In Natural Language Processing (NLP), source code is treated similarly to natural language, as noted by Contreras et al., "we convert samples into tokens and apply a word embedding model" [8]. Techniques such as tokenization, n-grams, TF-IDF, and WordbVec are commonly used to extract linguistic and structural features from code for model training [3] [8].

Transformer-based models like CodeBERT and GraphCodeBERT have demonstrated strong performance in this field [3] [4]. Jiang et al. report that "the performance of all pre-trained models significantly improves when combined with DFEPT," citing CodeBERT and GraphCodeBERT among the best-performing models [4].

Neuro-symbolic approaches, such as IRIS, combine large language models (LLMs) with static analysis to enable whole-repository reasoning and specification inference. According to Li et al., "IRIS combines static taint analysis with LLMs to infer specifications," thereby enabling contextual reasoning that reduces manual effort [9]. While LLMs alone often lack deep project-wide understanding, their combination with static analysis offers a promising avenue for improved security analysis: "Combining LLMs with static analysis boosts performance and correctness" [9].

Traditional static analysis techniques also struggle with the dynamic features of modern Java frameworks like Spring, particularly Dependency Injection (DI) and Aspect-Oriented Programming (AOP). Jasmine addresses this limitation by improving call graph completeness, enabling better analysis of Spring-based applications. As stated in the Jasmine paper, "static analysis tools have difficulty supporting core Spring technologies such as DI and AOP," which motivates the integration of AI-enhanced approaches [7].

A. Leveraging CWE in AI-driven Detection

AI models are trained on datasets comprising vulnerable code, annotated with CWE (Common Weakness Enumeration) IDs, to classify new code snippets [2] [5] [9]. By focusing on CWEs, researchers can develop AI models specifically tailored to detect classes of vulnerabilities, potentially improving detection accuracy and reducing false positives [4] [9]. Research indicates that training separate classifiers for each CWE significantly improves vulnerability detection accuracy, as these specialized classifiers effectively capture the unique "characteristics, code semantics, and patterns" associated with each vulnerability category, while treating all vulnerabilities as a single category "may oversimplify the problem, as it fails to capture the nuances and context specific to each CWE" [5].

Recent studies are actively exploring the effectiveness of AI models in detecting and classifying diverse vulnerability types, prominently including CWE-119 (Improper Restriction of Operations within the Bounds of a Memory Buffer), CWE-20 (Improper Input Validation), CWE-190 (Integer Overflow), and CWE-22 (Path Traversal). Harzevili et al. note that among various studies surveyed, "CWE-119, CWE-20, and CWE-190 are the most frequent ones" [3] while Li et al. specifically address the detection of "CWE-22: Path-Traversal" vulnerabilities through combined static analysis and large language models [9].

B. Vulnerable code datasets

The datasets for training these AI models originate from reputable sources such as the National Vulnerability Database (NVD) and the Software Assurance Reference Dataset (SARD), containing manually annotated CWE IDs for labeling. Rajapaksha et al. emphasize that their dataset creation involved "a web crawler... to collect a real-world dataset of 600,000 source code samples, which were annotated using static analyzers" [2]. High-quality datasets include PrimeVul, Juliet Test Suite, OWASP Benchmark [3] [9]. The PrimeVul dataset implements rigorous data de-duplication methods and evaluation criteria, highlighting the limitations of traditional ML standards [4]. In comparison, the Juliet Test Suite and OWASP Benchmark can assess static analysis tools' performance [3] [9].

Such datasets, which hold information on real-world code and experts' annotations, are crucial since "the lack of large and high-quality datasets for vulnerability detection is still a major obstacle" in practical vulnerability detection scenarios [5]. For instance, IRIS is evaluated on CWE-Bench-Java, "a dataset of manually validated security vulnerabilities" specifically designed around several common CWE categories [9].

IV. Security Detection in Java and Spring Framework

The Spring framework, being highly popular for enterprise Java applications, presents specific challenges for static analysis due to its dynamic features such as Dependency Injection (DI) and Aspect-Oriented Programming (AOP). These features obscure control and data flow, making traditional static analysis tools less effective [12].

To address this, Jasmine, a static analysis framework, extends the capabilities of Soot by enhancing call graph completeness for Spring applications. This is achieved by simulating Spring's dynamic behavior at the intermediate representation (IR) level [7]. Beyond static analysis, recent studies have adopted measurement-based approaches to identify security anti-patterns and insecure defaults in the configuration of Spring Security in enterprise Java applications, a module commonly used for authentication and authorization in Spring applications. Findings reveal recurring security anti-patterns, including disabled CSRF protection, hardcoded secrets, and lifelong access tokens [6]. While current research does not explicitly employ AI to detect such misconfigurations, the recurring nature and pattern-based signatures of these issues suggest potential for AI-assisted detection mechanisms in future work. Spring Boot and Java applications frequently face vulnerabilities related to CWE-79 (Cross-Site Scripting), CWE-89 (SQL Injection), CWE-94 (Code Injection), CWE-287 (Improper Authentication), CWE-285 (Improper Authorization), and CWE-611 (XML External Entity Injection) [3] [5]. The coverage and effectiveness of static analysis tools designed specifically for Java and Spring Boot applications remain limited. Research highlights persistent gaps in detecting framework-specific configurations, with Spring Boot's defaults often contributing to insecure deployments [5].

TABLE I
Comparison of AI-driven Vulnerability Detection Methods

Method	Strengths	Accuracy Metrics
JFinder	Quad self-attention	Accuracy: 97% (CWE)
IRIS	Neuro-symbolic	+28 vulnerabilities vs. CodeQL
DFEPT	Data flow embedding	Accuracy: 64.97% (Devign)
CWE-specific	CWE-focused accuracy	Varies by CWE type
Jasmine	Improved call graphs	Enhanced completeness

V. Summary and Discussions

This comprehensive review critically evaluates current methodologies in AI-driven vulnerability detection, specifically those employing static and data flow analyses aligned with the Common Weakness Enumeration taxonomy. Recent innovations, including deep learning architectures, graph-based representations, and neuro-symbolic methods, have notably increased detection accuracy; however, several significant limitations remain. Key among these limitations is the insufficient context-awareness exhibited by existing models, particularly when interpreting complex, framework-specific semantics inherent to Java and Spring Boot applications. A major impediment is the lack of comprehensive, labeled datasets specifically curated for Spring Boot environments, restricting effective model training, generalization capabilities, and accurate benchmarking in enterprise contexts. Accurately mapping real-world code to specific CWE identifiers remains challenging, resulting in elevated false-positive rates and incomplete detection coverage. Table I summarizes key AI-driven vulnerability detection methods discussed, highlighting their strengths and measured accuracy.

Despite their promising results, deep learning-based approaches such as CNNs, RNNs, and transformer models often face critical scalability issues in real-world software development environments. The computational overhead required to train and deploy these sophisticated models on large codebases or in continuous integration pipelines remains substantial, limiting their practical applicability in resource-constrained or real-time scenarios. Additionally, deep learning models present challenges for developers attempting to understand model predictions or debug false-positive detections. Future research should address these issues by focusing on lighter, more inter-

pretable models, or by hybridizing deep learning methods with traditional rule-based techniques to balance computational efficiency with detection accuracy. These findings underscore the necessity for developing a cohesive hybrid framework that unifies static code analysis, data flow analysis, and machine learning techniques, rooted in the CWE taxonomy for structured vulnerability categorization. Addressing these research gaps requires an innovative AI-driven model explicitly designed for Java and Spring Boot applications. This model would integrate CWE-specific classifiers, advanced semantic comprehension, and dynamic configuration awareness, thus substantially reducing false positives and enhancing detection accuracy. By closely aligning detection methodologies with the structural and behavioral nuances of enterprise Java applications, future research efforts promise significant improvements in the robustness, contextual relevance, and practical utility of automated vulnerability detection systems.

REFERENCES

[1] MITRE, "Common Weakness Enumeration (CWE) - A Community-Developed List of Software and Hardware Weakness Types," [Online]. Available: https://cwe.mitre.org, Accessed: Mar. 31, 2025.

[2] S. Rajapaksha, J. Senanayake, H. Kalutarage, and M. O. Al-Kadri, "Enhancing Security Assurance in Software Development: AI-Based Vulnerable Code Detection with Static Analysis," in Computer Security. ESORICS 2023 International Workshops, Cham: Springer, 2024, pp. 341–356.

[3] N. S. Harzevili, J. A. Ansari, H. Kazmi, and B. R. Shrestha, "A Systematic Literature Review on Automated Software Vulnerability Detection Using Machine Learning," ACM Computing Surveys, vol. 58, no. 1, pp. 1–38, Jan. 2025.

[4] Z. Jiang et al., "DFEPT: Data Flow Embedding for Enhancing Pre-Trained Model Based Vulnerability Detection," in Proc. 15th Asia-Pacific Symposium on Internetware, Macau, Jul. 2024.

[5] S. Al Atiiq, C. Gehrmann, K. Dahlén, and K. Khalil, "From Generalist to Specialist: Exploring CWE-Specific Vulnerability Detection," arXiv preprint arXiv:2408.02329, Aug. 2024.

[6] M. Islam, S. Rahaman, N. Meng, B. Hassanshahi, and P. Krishnan, "Coding Practices and Recommendations of Spring Security for Enterprise Applications," in Proc. IEEE Secure Development Conference (SecDev), 2020, pp. 49–56.

[7] M. Chen, T. Tu, H. Zhang, Q. Wen, and W. Wang, "Jasmine: A Static Analysis Framework for Spring Core Technologies," in Proc. 37th IEEE/ACM Int. Conf. Automated Software Engineering (ASE), Rochester, MI, USA, Oct. 2022, pp. 1–13.

[8] C. Contreras et al., "Multiclass Classification of Software Vulnerabilities with Deep Learning," in Proc. 15th Int. Conf. Machine Learning and Computing (ICMLC), Zhuhai, China, 2023, pp. 134–140.

[9] Z. Li, S. Dutta, and M. Naik, "LLM-Assisted Static Analysis for Detecting Security Vulnerabilities," arXiv preprint arXiv:2405.17238, Nov. 2024.

[10] L. Lambers, L. Sakizloglou, T. Khakharova, and F. Orejas, "Taint Analysis for Graph APIs Focusing on Broken Access Control," arXiv preprint arXiv:2501.08947, Jan. 2025.

[11] J. Wang, Z. Huang, H. Xiao, and Y. Xiao, "JFinder: A Novel Architecture for Java Vulnerability Identification Based Quad Self-Attention and Pre-training Mechanism," arXiv preprint arXiv:2307.15915, Jul. 2023.

[12] Pivotal Software, Inc., "Spring Framework," Spring.io, [Online]. Available: https://spring.io/projects/spring-framework. Accessed: Mar. 31, 2025.

AI Incorporated Job Matching System for Creating Career Opportunities

1st Narcis Nacev
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
narcis.karanfilov@gmail.com

2nd Nikola Kostadinov
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
nikolakostadinov2004@gmail.com

3rd Viktor Denkovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
viktor.denkovski@uacs.edu.mk

4th Marija Stankova Medarovska
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
marija.s.medarovska@uacs.edu.mk

***Abstract*—Besides academic achievements and gaining technical competence, looking at and successfully browsing the employment market presents one of the biggest stumbling blocks not just for information technology (IT) graduates, but overall, for all candidates pursuing a job and career opportunity. With so many available platforms offering and publishing job opportunities, it becomes frustrating and difficult to keep up and find the right employment. This paper introduces NextMatch, an artificial intelligence (AI) incorporated system with data analysing functionality that compares job postings from LinkedIn to enhance the efficiency of the job search process. Using a data-driven approach, NextMatch determines compatibility scores and generates tailored job recommendations, indicating the likelihood of being accepted for each position. The results demonstrate the potential of AI-powered systems to reduce job search time and increase placement success by offering personalized job recommendations. This study contributes to the ongoing development of AI-based job recommendation systems, emphasizing the integration of personalized matching, advanced algorithms, and enhanced transparency to ensure better alignment between job seekers and employers.**

Keywords—job matching, AI-driven recommendations, LinkedIn integration, career opportunities, compatibility scoring

I. Introduction

One of the greatest challenges for any individual, not just recent graduates, is searching for a new job opportunity, a process often accompanied by difficulties and obstacles in job scouting and career planning [1]. Although there are more and more employment platforms, such as Indeed and LinkedIn, which are designed to help with the job search, many recent graduates find it difficult to search for positions that correspond to their skills and goals for the future. The conventional way of job seeking often relies on manual filtering and keyword searching which can be time-consuming and ineffective [2]. Bearing in mind the limitation in capacity to gather contextual information, the traditional algorithms for job-matching usually do not provide ideal recommendations as they mostly rely on the methods mentioned above [3]. Furthermore, employers often prioritize soft skills and experience that might not be well presented in traditional job descriptions and resumes [4]. These elements highlight the requirements for a more efficient way to connect graduates with adequate employment corresponding to their goals, experiences, and qualifications. Hence, the introduction of automated job matching systems was developed to fill in this gap and ease the process by using data-driven methods and AI to improve and speed up the procedure [5].

This paper has a goal to examine the integration and efficiency of NextMatch in improving the job searching experience for IT professionals and computer science graduates. The system uses a data-driven approach that matches relevant employment opportunities with graduates by taking and analysing their resumes and comparing them with job postings on the LinkedIn platform. For the system to offer insights into the possibility of acquiring a position based on their profiles and to increase employment search efficiency, the system calculates a compatibility score and enables users to see their personalised recommendations. By implementing more precise analysis based on skills, this data-driven strategy tries to optimize the techniques for job-searching and lower the time spent on job applications. The presented approach could accelerate career transitions and provide graduates with decisions that are well-informed by filling in the missing parts in the current job-matching techniques.

As such, this paper is divided into six sections in which the first introduces the goals of the research. The second section presents some related corresponding work done to this research. The third section presents the materials and methods used for the development process, operations and functions. The fourth section presents the gathered results of the study. The discussion of the results is presented in section five. Finally, section six concludes the paper.

II. Related Work

In recent years, the increasing use of job recommender systems (JRS) has significantly enhanced the job search process by aligning candidates' skills with suitable employment opportunities. By implementing ML models that evaluate talents, contextual data, and job requirements, AI-driven job recommendation systems try to improve on the negatives of traditional keyword-based searches [6]. To raise the accuracy of recommending matching jobs, these systems use NLP techniques to extract and group relevant data from CVs and job descriptions [7].

AI-driven solutions for job searches that improve candidate-to-employer matching have been suggested in various studies. A comprehensive survey done by de Ruijt, and Bhulai [8] reviews various JRS methodologies, including content-based, collaborative filtering, and hybrid approaches, discussing their respective advantages and challenges. Further advancements are highlighted in Zheng et al. [9], which introduces a generative job recommendation model based on large language models (LLMs). This approach aims to provide personalized job suggestions without relying solely on existing job databases, offering a more comprehensive career advisory system. Additionally, Preetam and Sadaphal [10] proposed JobRecoGPT, an explainable job recommendation system utilizing LLMs to capture refined information from job descriptions and resumes, addressing the limitations of traditional methods. Explainability in AI models is yet another crucial segment of job matching. Understanding the rationale behind job recommendations enhances transparency and allows job seekers to make more informed career decisions. Recent advancements in explainable AI have significantly improved the interpretability of AI-driven recommendations. This increased transparency not only helps job seekers comprehend the alignment between their profiles and job opportunities but also boosts their confidence in the system, ultimately facilitating more strategic job applications.

These studies collectively demonstrate the evolving landscape of job recommendation systems, emphasizing the importance of transparency, fairness, and the incorporation of advanced language models to enhance the matching process between job seekers and employers. Addressing a critical challenge in the job market by joining skill-based scoring, explainability, and personalized recommendation, the recommended system improves the effectiveness of job-seeking.

III. Materials And Methods

A. Implementation Details

Searching for job and internship opportunities can be an overwhelming and time-consuming process. Candidates often need to go through extensive job descriptions and requirements to identify relevant positions. To address this challenge, an automated system, NextMatch, was developed to streamline job searching and matching. This system integrates various technologies to efficiently gather, process, and present job-related information to users.

The system's backend is developed using Python [11], and the Django REST framework is employed to build and manage API endpoints, enabling communication between the backend and frontend [12]. Additionally, an SQL-based database (PostgreSQL/SQLite) is used to store structured information about job postings, user profiles, and match percentages [13, 14]. To ensure proper functionality, Postman is utilized for testing API endpoints [15].

To automate job data retrieval, ScrapingDog API is used to extract job listings from LinkedIn [16]. The frontend is built using React [17], a JavaScript framework that dynamically presents job recommendations to users [18]. Node.js serves as the runtime environment for handling frontend development [19], while npm (Node Package Manager) facilitates dependency management [20]. Chakra UI is integrated to provide a modern and accessible design for a better user experience [21]. TypeScript was used for frontend development, enhancing type safety and maintainability [22]. The system's job-matching process involves extracting and analyzing CV content using PDFPlumber [23], where the extracted information is then compared with job descriptions stored in the database to generate compatibility scores. Fig. 1 presents the overall architecture design of the system.

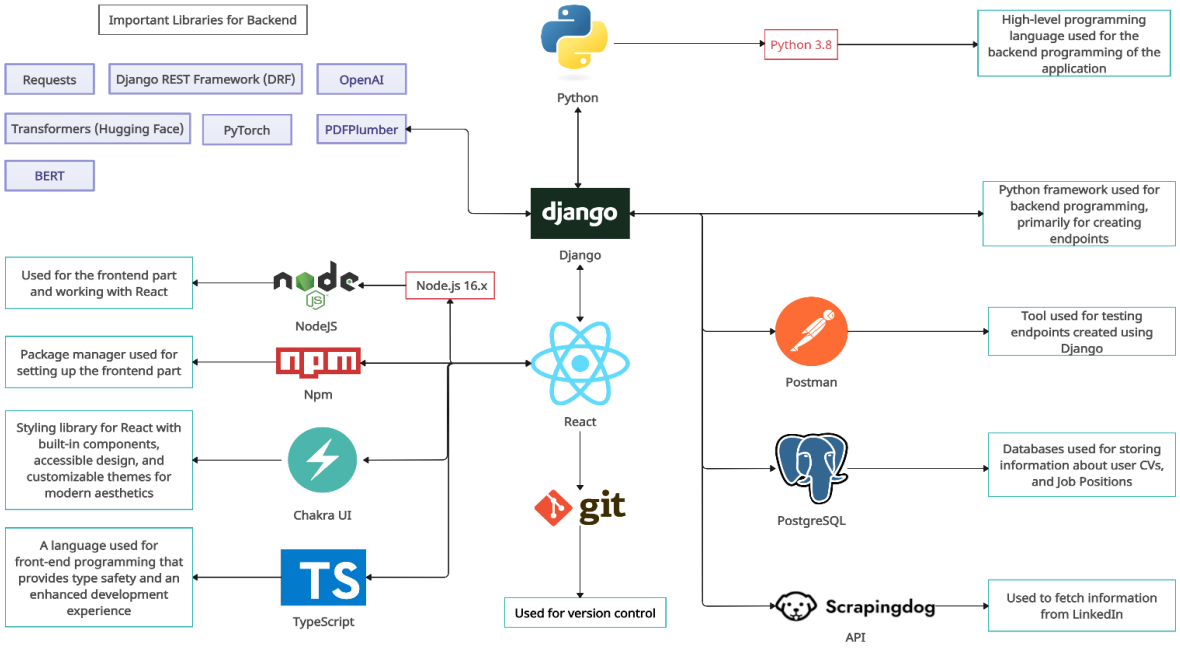


Fig. 1. The overall architectural design of the NextMatch system

In the initial phase of system interaction, the user is presented with a registration form that requests essential personal details and a resume from the user (Fig. 2).

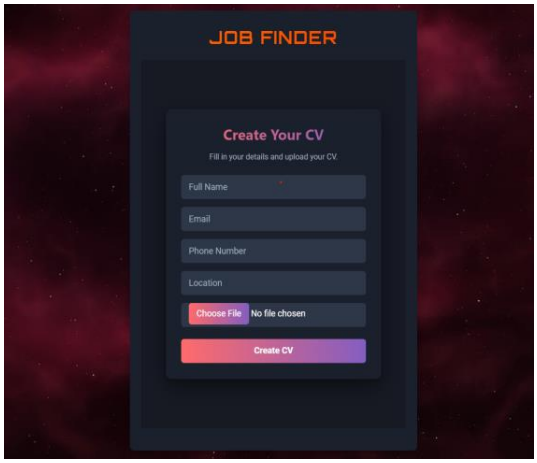


Fig. 2. Initial CV creation starting page

From there, the CV is passed to the backend, where a function extracts the provided sections and saves them in the PostgreSQL database model for CV, as shown in Fig. 3.

Cv	
full_name	CharField
email	EmailField
phone_number	CharField
location	CharField
skills	TextField
education	TextFiled
experience	TextField
cv_file	FileField
uploaded_at	DateTimeField

Fig. 3. Saved sections of the CV model in PostgreSQL

After that, it fetches the job information and places it in the JobPosition model, as shown in Fig. 4.

JobPosition	
job_position	CharField
company_name	CharField
job_location	CharField
job_posting_date	CharField
job_description	TextField
job_link	URLField
seniority_level	CharField
employment_type	CharField
job_function	CharField
industries	CharField
job_apply_link	URLField

Fig. 4. Saved sections of the job model in PostgreSQL

After the functions that fetch the jobs and their details and complete their operation, it emits an HTTP of 201 meaning it was a success to the frontend which then prompts a button to pop up on the screen that says, “Find Jobs”. Upon clicking the button, the job recommendations shown in Fig. 5 as a sample are presented.

B. *Recommender System*

The recommender system serves as the core component of the NextMatch application, enabling personalized job recommendations by matching candidate resumes with relevant job postings. As a content-based recommender system, it analyses textual features extracted from both resumes and job descriptions to determine compatibility. Two methods were evaluated for this purpose:

1. Term Frequency-Inverse Document Frequency (TF-IDF) with cosine similarity; and
2. Bidirectional Encoder Representations from Transformers (BERT) embeddings.

TF-IDF is a statistical measure used in information retrieval to assess the importance of words in a document relative to a corpus. It assigns higher weights to terms that appear frequently in a document but are rare across the corpus, effectively filtering out common words while retaining meaningful ones [24]. Combined with cosine similarity, which measures the angle between document vectors in high-dimensional space, TF-IDF enables efficient yet interpretable relevance scoring between resumes and job descriptions [25]. However, TF-IDF is limited in capturing semantic meaning since it treats words as independent features. In contrast, BERT embeddings provide contextualized word representations, allowing the system to capture deeper semantic relationships between job postings and candidate resumes [26]. Unlike TF-IDF, BERT embeddings consider word meaning based on surrounding context, enabling a deeper understanding of the text.

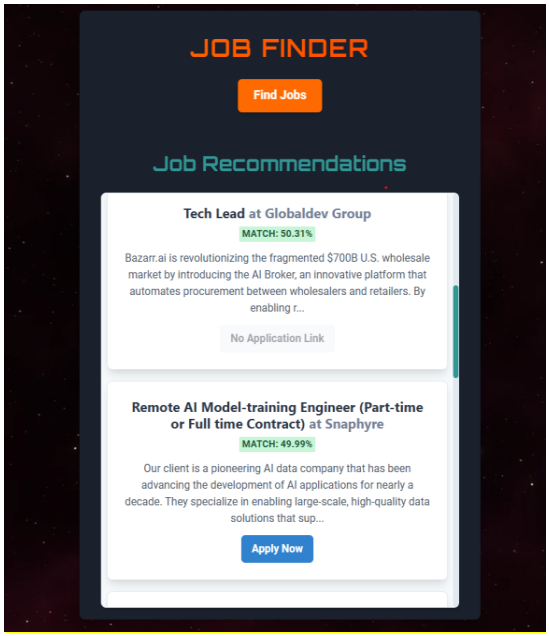


Fig. 5. Sample result of matched jobs by using the system NextMatch

C. *Data*

To evaluate the performance of the JRS, a dataset of 700 job postings was sourced from LinkedIn, covering a diverse range of IT job roles on a national market level, all published in English language. Additionally, 120 anonymized IT professional resumes were obtained from Kaggle/Hugging Face [30, 31] in a PDF format. These resumes were collected by scraping publicly available resume examples from an online recruiting platform present in the United States (US). To comply with ethical and privacy standards, all personally identifiable information (PII) are removed from the dataset. Since no labels are available, GPT-4o was employed as an automated evaluator to assess the quality of the recommendations, through the OpenAI API [27]. For each candidate's resume, the top three job recommendations from both methods were rated on a scale of 1-10, where 1 represents the lowest match and 10 is the highest. This approach aligns with recent research that highlights the

potential of large language models (LLMs) as evaluators for subjective and qualitative tasks. Studies such as Zheng et al. [28] suggest that LLM-based evaluators perform similarly to human evaluators in controlled assessments, in tasks where explicit ground-truth labels are not available. The prompt used for generating the evaluation is provided in Fig 6.

```
# Objective:

You are an impartial AI judge for NextMatch, a job recommender system. Score how well
a candidate's resume matches job vacancies from Method A and Method B on a 1-100
scale (1 = No match, 100 = Perfect fit).

# Input:

{ "candidate_resume": "{resume_text}",
 "method_a_jobs": ["{method_a_job_1}", "{method_a_job_2}", "{method_a_job_3}"],
 "method_b_jobs": ["{method_b_job_1}", "{method_b_job_2}", "{method_b_job_3}"]}

# Output:

{ "method_a_scores": [X, X, X],
 "method_b_scores": [X, X, X] }

Where the score X is rated on a scale from 1 to 100 based on skills, experience, education,
and role relevance. Ensure fair and unbiased scoring.
```

Fig. 6. Prompt used for querying ChatGPT to impartially evaluate the top 3 highest recommendations provided by: (a) TF_IDF and cosine similarity and (b) BERT embeddings

IV. Results

A comparative evaluation of job recommendation methods using Term Frequency-Inverse Document Frequency (TF-IDF) with cosine similarity and BERT embeddings revealed no statistically significant difference in the overall recommendation quality, as validated by a Wilcoxon Signed-Rank Test ($p > 0.05$) [32]. Nevertheless, an analysis of recommendation overlap indicated that only 6.79% of job postings were commonly recommended by both methods within the top three rankings for the same candidate. This low overlap suggests that the two approaches may capture distinct features of job-candidate matching. TF-IDF, which assigns weights to terms based on their frequency in a document relative to their inverse frequency across a corpus, performs well in keyword-based matching tasks. In contrast, BERT, uses transformer-based deep contextual embeddings, and captures semantic similarities by modeling bidirectional relationships between words. However, this increased sensitivity to context could result in greater variability in recommendations, as BERT may interpret minor differences in phrasing or context as significant, potentially affecting consistency across similar inputs.

An examination of the top recommendation for each candidate found that BERT embeddings achieved a higher mean GPT-4o evaluation score of 0.61 (SD = 0.23) compared to TF-IDF with Cosine Similarity 0.58, but also with a greater variability (SD = 0.20). The analysis of low-scoring and high-scoring CVs reveals key differences in job recommendation effectiveness. Candidates whose top recommendations from both methods scored below 0.25 often had limited, vague or highly specialized skills and experience, making it difficult for the system to find strong matches. In contrast, candidates with top recommendations above 0.75 had well-defined, in-demand skills and detailed descriptions of experience. Overall, while neither method consistently outperformed the other, their distinct strengths suggest that a hybrid approach could provide more robust and reliable job recommendations.

V. Discussion

By integrating advanced AI techniques into the NextMatch system with a user-friendly design, the platform helps address an important challenge in career development for computer science graduates and other candidates looking for employment opportunities by using artificial intelligence to simplify and enhance the job matching process. This approach aligns with recent advancements in job recommender systems (JRS) highlighted by existing research:

- AI-Driven Matching: The system demonstrates the potential of using ML techniques to match candidate profiles with job requirements more effectively than traditional job search methods [29];
- Integration of Advanced Technologies: By combining multiple technologies (Django, React, ScrapingDog API, PDFPlumber), NextMatch showcases a comprehensive approach to job recommendation that goes beyond conventional platforms;
- Personalization: The compatibility scoring mechanism provides graduates with tailored job recommendations, reducing the overwhelming nature of job searching and increasing efficiency (Fig. 7);
- Comparative Analysis: This approach shares similarities with recent studies [9] and [10], which also emphasize AI-driven, personalized job recommendations.

Unlike traditional methods, NextMatch:

- Extracts and analyzes CV information dynamically;
- Provides percentage-based compatibility scores;
- Offers a possibility for real-time job matching across multiple platforms.

The system, although very functional, comes with its current limitations, such as: expanding job source integrations beyond LinkedIn, enhancing the matching algorithm's precision, and implementing more sophisticated ML models for recommendations.

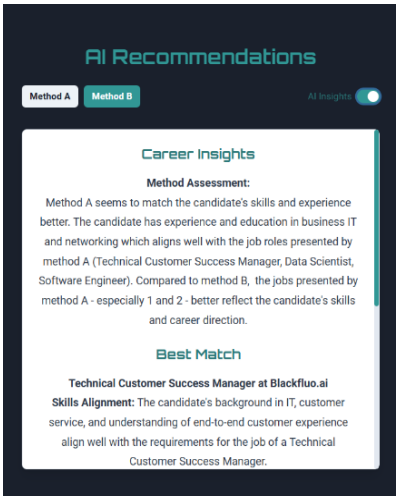


Fig. 7. AI recommendations for career insights from NextMatch

Having to consider all the above, the practical implications of the system NextMatch represent a promising solution for recent graduates navigating complex job markets, potentially reducing job search time and increasing placement success rates. By integrating advanced AI techniques with user-friendly design, the platform addresses a significant obstacle point in career development for emerging professionals in the technology sector.

CONCLUSION

In conclusion, the NextMatch system presents a promising solution to the challenges job seekers, particularly recent graduates, face in navigating the complex and often overwhelming job market. This research highlights its potential to streamline and improve the job-matching process for students and IT professionals alike. Rather than enduring the time-consuming, repetitive task of applying to countless positions of varying relevance, NextMatch offers a more efficient and targeted approach. By leveraging advanced algorithms, data-driven insights, and a user-centric design, the system matches candidates with opportunities that align closely with their skills, preferences, and career aspirations. This reduces the manual effort required in job hunting, enabling users to focus on refining their skills and preparing for roles best suited to their strengths.

However, ensuring accessibility for all users while addressing concerns such as data privacy and platform reliability remains essential and needs further exploration. Current job-matching systems, including NextMatch, face challenges like inaccurate job postings, potential biases in AI models, and difficulties in assessing users' soft skills. Overcoming these limitations requires ongoing refinement and innovation.

Looking ahead, future work could dive deeper into improving the personalization and efficiency of the platform. Expanding the scope to include global markets, remote work opportunities, and skill-building resources could make the platform even more robust. Future directions of this work could focus on performing a more exhaustive evaluation considering a greater number of methods and data as well. Also, the aim was to supplement GPT-4o's subjective scoring with a user study involving IT and HR professionals to assess perceived relevance and satisfaction. Although ground truth labels were unavailable, proxy metrics offered a practical evaluation framework, laying a solid foundation for future investigations.

REFERENCES

.

[1] Wilson, M. J., Wood, L., Solomonides, I., Dixon, P., and Goos, M. "Navigating the Career Transition from Industry to Academia." *Industry and Higher Education*, vol. 28, no. 1, Feb. 2014, pp. 5–13, https://doi.org/10.5367/ihe.2014.0189.

[2] Sivabalan, L., Yazdanifard, R., and Ismail, N. H. "How to Transform the Traditional Way of Recruitment into Online System." *International Business Research*, vol. 7, no. 3, Feb. 2014, https://doi.org/10.5539/ibr.v7n3p178.

[3] Goda, S., Hayashi, Y., and Saito, Y. (2024). A Best-of-Both Approach to Improve Match Predictions and Reciprocal Recommendations for Job Search. *ArXiv*, *abs/2409.10992*.

[4] World Economic Forum. "Future of Jobs Report 2023." World Economic Forum, World Economic Forum, May 2023, www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf.

[5] "Artificial Intelligence and Labour Market Matching." *OECD Social, Employment and Migration Working Papers*, Jan. 2023, https://doi.org/10.1787/2b440821-en.

[6] Thali, R., Mayekar, S., More, S., Barhate, S., and Selvan, S. Survey on Job Recommendation Systems Using Machine Learning. Mar. 2023, pp. 453–57, https://doi.org/10.1109/icidca56705.2023.10100122. Accessed 20 Feb. 2025.

[7] Supriyono, N., Prasetya Wibawa, A., Suyono, N., and Kurniawan, F. "Advancements in Natural Language Processing: Implications, Challenges, and Future Directions." Telematics and Informatics Reports, Elsevier BV, Nov. 2024, pp. 100173–73.

[8] de Ruijt, C., and Bhulai, S. "Job Recommender Systems: A Review." *ArXiv.org*, 26 Nov. 2021, https://doi.org/10.48550/arXiv.2111.13576. Accessed 30 May 2023.

[9] Zheng, Z., Qiu, Z., Hu, X., Wu, L., Zhu, H., and Xiong, H. "Generative Job Recommendations with Large Language Model." *ArXiv.org*, 5 July 2023, https://doi.org/10.48550/arXiv.2307.02157.

[10] Preetam, G., and Sadaphal, V. "JobRecoGPT -- Explainable Job Recommendations Using LLMs." *ArXiv (Cornell University)*, Cornell University, Sept. 2023, https://doi.org/10.48550/arxiv.2309.11805. Accessed 15 Mar. 2024.

[11] PYTHON. "Python." *Python.org*, Python.org, 29 May 2019, www.python.org/.

[12] Christie, T. "Home - Django REST Framework." *Django-Rest-Framework.org*, 2011, www.django-rest-framework.org/.

[13] SQLite. "SQLite Home Page." *Www.sqlite.org*, 7 Dec. 2024, www.sqlite.org/.

[14] The PostgreSQL Global Development Group. "PostgreSQL: The World's Most Advanced Open Source Database." *Postgresql.org*, 2019, www.postgresql.org/.

[15] Postman. "Postman | the Collaboration Platform for API Development." *Postman*, 2021, www.postman.com/.

[16] "Scrapingdog: Best Web Scraping API." *Scrapingdog*, 7 Feb. 2025, www.scrapingdog.com/?deal=jorge20&gad_source=1&gclid=Cj0KCQiA-5a9BhCBARIsACwMkJ6VKvdHRDlllZA4P7U-yPESKg7FN-on-OoDNGEkYLk9C5EyyvAQmKsaAqEEEALw_wcB. Accessed 7 Feb. 2025.

[17] Meta Open Source. "React." *React.dev*, 2024, react.dev/.

[18] *Yarnpkg.com*, 2019, yarnpkg.com/.

[19] Node.js. "Node.js." *Node.js*, 2023, nodejs.org/en.

[20] "Frontend." *Npm*, 9 Feb. 2019, www.npmjs.com/package/frontend. Accessed 7 Feb. 2025.

[21] "Chakra UI - a Simple, Modular and Accessible Component Library That Gives You the Building Blocks You Need to Build Your React Applications." *Chakra UI: Simple, Modular and Accessible UI Components for Your React Applications.*, chakra-ui.com/.

[22] Microsoft. "TypeScript - JavaScript That Scales." *Typescriptlang.org*, 2015, www.typescriptlang.org/.

[23] "PDFPlumber | LangChain." *Langchain.com*, 2020, python.langchain.com/docs/integrations/document_loaders/pdfplumber/. Accessed 7 Feb. 2025.

[24] Ramos, J., 2003, December. Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48).

[25] Schütze, H., Manning, C.D. and Raghavan, P., 2008. Introduction to information retrieval (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

[26] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[27] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[28] Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36, pp.46595-46623.

[29] Rahman, A., Paul, A., Rahman, F., Akter, J., Sultana, Z., and Rahman, M. "Appropriate Job Selection Using Machine Learning Techniques." *Research Square (Research Square)*, Research Square (United States), July 2023, https://doi.org/10.21203/rs.3.rs-3164137/v1. Accessed 9 Jan. 2024.

[30] https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset

[31] Mohr, D.L., 2010. Statistical methods. Academic Press.

Ciphertext-Only Attacks Using Large Language Models

Marija Taneska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
marija.taneska@finki.ukim.mk

Ikechukwu John Chukwu
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
ikechukwu.chukwu@students.finki.ukim.mk

***Abstract*—Ciphertext-only attacks, a fundamental cryptanalytic technique, assume an attacker possesses only ciphertext. This research investigates the potential of Large Language Models (LLMs) in performing ciphertext-only attacks against the zigzag cipher. We explore the feasibility of LLM-driven cryptanalysis, acknowledging that recent advancements enable LLMs to tackle previously challenging tasks, including cryptographic attacks. Our study designs and implements an LLM-based system for this purpose. Our findings demonstrate near-perfect attack success when the key length is less than 4. However, we also observed inconsistencies in LLM performance, necessitating further investigation. These results highlight the growing relevance of LLMs in cryptanalysis and underscore the need for increased awareness within the cryptographic community. Furthermore, this study suggests future directions for research regarding the evolving landscape of cryptanalysis in the context of advanced AI.**

***Index Terms*—Cryptography, Cryptanalysis, Classical ciphers, Rail Fence, LLM, Large Language Models, ChatGPT, Gemini, Ciphertext-only attacks, COA.**

I. Introduction

Classical cryptography provided the foundation for secret communications for centuries based on the substitution and transposition of plaintext. Although these traditional methods are almost obsolete in modern security, they provide fertile ground for exploring the foundations of cryptanalysis techniques, such as the ciphertext-only attack (COA). This is considered a minimal-information attack model and a stringent test for cryptographic strength and the attacker's inference capabilities.

The Rail Fence cipher is a classical transposition cipher, which encrypts a given message by placing the characters into a series of "rails" or rows in a zig-zag fashion and then reading those rows one at a time to produce the ciphertext. While the number of rails used acts as the encryption key, this is unknown when attacking ciphertext-only.

Large language models (LLMs) have demonstrated significant capabilities in a wide range of natural language processing tasks, including but not limited to code generation, translation, and reasoning. While the ethical implications of these models have been widely discussed, less attention has been given to their potential to deduce what original plaintexts correspond to ciphered messages from statistical and linguistic tendencies alone.

This research aims to compare the output of ChatGPT and Gemini towards recovering plaintext from Rail Fence encrypted words, assuming no information apart from the type of cipher used. We want to see if these models can reverse-engineer the original message from ciphertext alone by recognizing letter patterns and inferring relationships based on probabilities.

The remainder of this paper is organized as follows. In section II, we review studies that have conducted some form of LLMs-based cryptanalysis. Section III details our proposed methodology, experimental design, and analysis. The results are summarized in Section IV, where we also discuss plausible and suspicious statistical relationships. Finally, before concluding in Section VI, Section V discusses the findings and proposes relevant future research directions.

II. Related work

Several works on cryptanalysis focused on using large language models to attempt cryptographic attacks. For example, [1] evaluated deep learning approaches on the complete columnar transposition cipher for known plaintext and proposed an algorithm to this effect. They concluded and reported that their algorithm achieved an accuracy of 96% in predicting keys, possibly because it was developed using a GRU-LSTM-hybrid architecture known for handling complex sequences.

A focal point of these attacks using large language models (LLMs) involves exploiting generative adversarial networks (GANs) to understand classical ciphers. Early works studied successfully exploited classical ciphers using the hidden Markov model (HMM) [2], [3], whose results were compared with [4]. In [4], instead of using the HMM technique, they exploited the effectiveness of CipherGAN [5] for attacking classical ciphers. Their results revealed that CipherGAN could solve polyalphabetic Vigenère cipher with prior extensive data training to be kept in mind. Like other studies using deep learning on COA, [4], [6] found that plaintext images can be retrieved from ciphertext images that are not known in real-time, that optical scanning can be exploited by deep learning COA, and that it is impractical and limited by a need for a large data set.

Lastly, recent works [7], [8] explored the problem of cryptographic API misuse using LLMs such as ChatGPT and

Genimi. In contrast, other works studied how implementing attack models can be feasible using ChatGPT [9], [10], [11]. An interesting aspect of these works for our study is that they compared static analysis tools with LLMs and highlighted the accuracy of LLMs in detecting misuse. Strong suggestions from those works were further expressed on LLMs' promising ability and potential in attack and analysis.

The key difference in our work is that we do not study how crypto APIs are misused but are the first to conduct a combined analysis using two LLMs on classical ciphers. To the best of our knowledge, our work, in contrast to previous studies, strongly focuses on the statistical properties of the cryptanalysis computation on two LLMs. To achieve this, we proposed and developed a system.

III. Experimental Methodology

A. System Architecture

We implemented a web application based on the Flask framework to perform controlled ciphertext-only attack experiments. This system allows automated interaction with large language models (LLMs)—specifically, ChatGPT and Gemini—providing an interface to upload test datasets, trigger model evaluations, and download structured results.

1) User Input: Before running an experiment, the user must specify the following parameters:

- The **name of the cipher** used for encryption (e.g., Rail Fence)
- A text **file containing encrypted words**
- A text **file with the corresponding plaintext words** (used only for evaluation)
- The **number of candidate plaintext guesses** to be requested from the model per encrypted word
- The **choice of LLM model** to use for decryption: ChatGPT or Gemini

2) Model Interaction: A dynamic prompt is constructed for each encrypted word based on the selected cipher and the number of requested plaintext guesses. The prompt specifies the cipher (e.g., Rail Fence) and instructs the model to return a fixed number of valid English words without additional formatting, numbering, or commentary. This format is designed to standardize responses between ChatGPT and Gemini and minimize differences in prompt interpretation. During execution, the completed prompt is sent to the selected LLM: ChatGPT via the OpenAI API or Gemini, the Google Generative AI API.

3) Evaluation: The LLM outputs a free-text response that is parsed and assessed against the following criteria:

1) **Exact Match** — A case-insensitive string comparison of the exact parison with the anticipated cleartext.
2) **Letter Frequency Match** - a character-by-character comparison of outcomes of those distributions to verify whether the selected word and the original have the same letters in any order.
3) **Length Match** — A comparison to see if the correct guess word has an identical character count to the guessed word.

Each decryption attempt is recorded with metadata, including its position in the input, whether it was a successful match, the time to generate the response, and whether it satisfied the letter frequency and length criteria. The results are compiled into a CSV file, enabling batch analysis and direct comparison of decryption behavior across the test set.

B. Testing Methodology

The experiments were conducted exclusively on the Rail Fence cipher, a classical transposition cipher in which the number of rails used during encryption was intentionally omitted from the prompt, simulating a ciphertext-only attack (COA) scenario. The goal was to determine whether a large language model could successfully infer the original plaintext based solely on the encrypted input and general linguistic knowledge.

Both models were tested under identical conditions using the same encrypted dataset and parameter settings to ensure consistency. The experiments were conducted with the latest versions of each language model at the time of testing. Specifically, `chatgpt-4o-latest` was used via the OpenAI API, and `gemini-1.5-flash` was used via the Google Generative AI API. Both models were accessed with default settings without additional fine-tuning or customization. For each model response, every guessed word was evaluated according to the criteria described in the evaluation subsection: exact match, letter frequency match, and length match. The time to receive the model's response was also measured and logged.

The test set consisted of 60 encrypted words evenly distributed across different combinations of word lengths (3, 5, and 7) and Rail Fence key sizes (2, 3, and 4). ChatGPT and Gemini were prompted independently for each configuration to produce up to three candidate plaintext guesses per word, resulting in 180 total decryption attempts per model. The choice of Rail Fence key sizes was aligned with the constraints of the cipher and the goal of testing practical ciphertext complexity. Only key size two was used for short words (length 3), as larger keys would reproduce the original text or provide negligible encryption. For medium-length words (length 5), keys 2 and 3 were selected to create more substantial transposition patterns while preserving readability. For longer words (length 7), keys 2, 3, and 4 were used, allowing more profound zigzag transformations that increased cryptographic complexity without making the output too sparse. This configuration ensured that each word size was paired with a realistic and non-trivial cipher setup, enabling consistent, comparative evaluation of both models under ciphertext-only attack conditions. All evaluation data was compiled into a CSV file, allowing for post-processing, statistical analysis, and comparison of model performance across the entire test set.

IV. Results and Analysis

We first provide an overview of the collected data, followed by a detailed analysis of the models' attack attempts. This

includes a breakdown of success and failure rates, an evaluation of the effectiveness of each attack try, an investigation into unsuccessful decryption attempts, and a comparison of computational performance across scenarios.

A. Success and Failure Rates

We compared the ratio of successful to unsuccessful attacks for ChatGPT and Gemini, as shown in Fig. 1. The data reveals an apparent performance disparity between ChatGPT and Gemini. ChatGPT consistently achieved higher success rates for longer word lengths and larger key sizes. While demonstrating initial success with short inputs, Gemini experienced a significant performance decline as input complexity increased.

For example, ChatGPT achieved a 100% success rate for word length 3 with key size 2, compared to Gemini's 60%. Similarly, at word length five and key size 2, ChatGPT maintained 90% success, while Gemini dropped to 50%.

The impact of increasing word length was also more severe for Gemini. At word length seven and key size 2, ChatGPT achieved 70% success, whereas Gemini fell to just 20%.

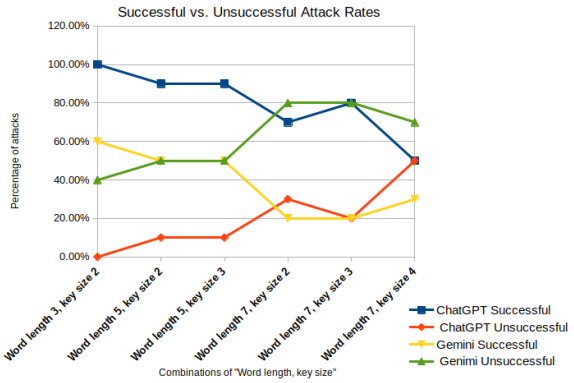


Fig. 1. Successful vs. Unsuccessful Attacks

This disparity underscores the varying capabilities of different LLM architectures in cryptanalytic tasks. The observed decline in performance for both models with increasing input complexity reinforces the challenge of deciphering complex ciphertext using LLMs.

ChatGPT showed greater resilience across input variations, while Gemini's effectiveness diminished rapidly as word length or key size increased.

Finally, these findings contribute to a comprehensive understanding of the potential and limitations of LLMs in ciphertext-only attacks. They also highlight the need for further research into the specific architectural features that influence LLM performance in cryptanalysis, as well as the importance for the cryptography community to acknowledge the potential threat LLMs pose in low-information attack scenarios.

B. Attack Attempt Analysis

We tracked success rates per try to understand which specific attempt (first, second, or third) was most likely to succeed. The results, shown in Fig. 2, show that the first guess is often the most effective, with significantly reduced success in subsequent attempts. This indicates that if LLMs can retrieve the correct plaintext, it will most likely appear at the top of their generated candidate list.

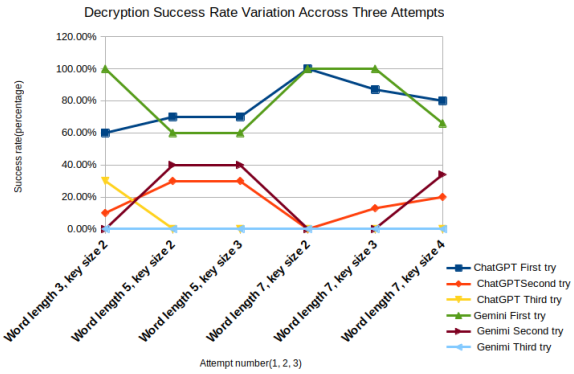


Fig. 2. LLM Attack Success Rates

For ChatGPT, the first try accounted for the most correct guesses across all configurations, with success rates ranging from 60% to 100%. In cases such as word length 7 with key size 2, 100% of ChatGPT's successful decryptions were on the first try. The second and third guesses showed only marginal contributions. For instance, at word length five and key size 2, the second guess contributed an additional 30%, while the third guess provided 0%. Gemini showed an even stronger dependency on the first try. In multiple test cases—including word lengths 3 and 7 with key sizes 2 or 3—all correct guesses occurred in the first attempt, with 0% success in the second and third attempts.

These findings reinforce the idea that when LLMs are successful at ciphertext-only decryption, the most probable output is typically placed first—possibly due to internal ranking mechanisms that favor high-confidence linguistic patterns.

C. Analysis of Unsuccessful Attacks

To further understand the behavior of LLMs in ciphertext-only attacks, we analyzed the characteristics of unsuccessful decryption attempts. Specifically, we examined two metrics: Length Match and Letter Frequency Match. The results are presented in Fig. 3.

Across all attacks, Gemini showed a higher length match rate than ChatGPT. For example, at word length three and key size 2, Gemini reached a 92% Length Match rate, compared to ChatGPT's 40%. However, Gemini performed significantly worse in terms of letter frequency matching. In the same attack, only 13% of its guesses matched letter frequencies, while ChatGPT achieved 25%. As word length increased, ChatGPT performance improved, especially with key size 3. At word length 5, and key size 3, it achieved 75% length match and 69% letter frequency match, its highest in the dataset. Conversely, Gemini showed a steep decline in both metrics as the input complexity grew. For word length 7 with key sizes

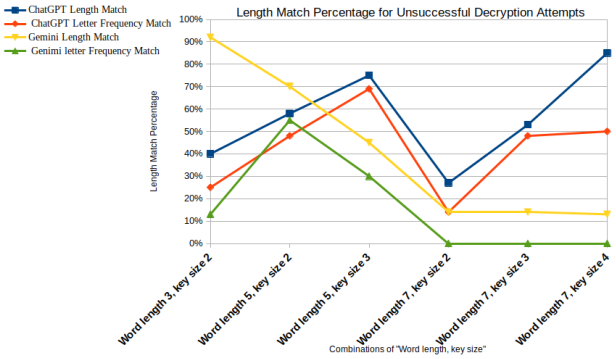


Fig. 3. Analysis of Unsuccessful Decryption Attempts

3 and 4, it dropped to 14% or less for the length match and 0% for the letter frequency match. These results suggest that ChatGPT produces more linguistically structured outputs even when incorrect, while Gemini's incorrect guesses deviate more drastically from the structure of the target word.

D. Computation Time Analysis

As illustrated in Fig. 4, the average computation time varied across different word lengths and key sizes for both LLMs. Overall, Gemini exhibited slightly faster computation times in most scenarios. However, the differences were not substantial, indicating that both LLMs can perform decryption attempts within a relatively short timeframe. It is worth noting that while Gemini was faster for the shorter word lengths, it was slightly slower for the most extended word length and key size combination. This could point towards how the LLMs handle complexity.

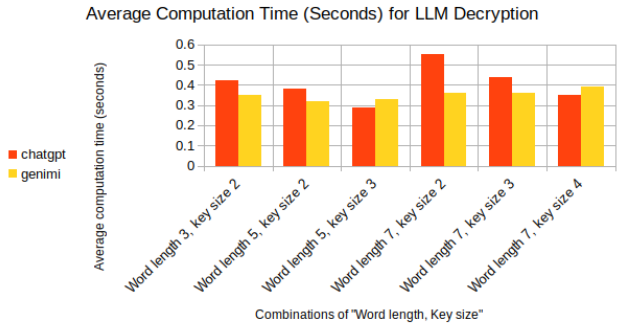


Fig. 4. Computation Time Across Word Lengths and Keys

This disparity underscores the varying capabilities of different LLM architectures in cryptanalytic tasks. The observed decline in performance for both models with increasing input complexity reinforces the challenge of deciphering complex ciphertext using LLMs.

Finally, these findings contribute to a comprehensive understanding of the potential and limitations of LLMs in ciphertext-only attacks. They also highlight the need for further research into the specific architectural features that influence LLM performance in cryptanalysis and the importance of the cryptography community acknowledging the potential threat LLMs pose in low-information attack scenarios.

V. Discussion and Future Work

Although this study examines one particular cipher and a limited number of language models, the framework was designed with future extensibility in mind. Moving forward, we plan to evaluate additional variants of ChatGPT and Gemini, including newer or fine-tuned versions, and incorporate alternative large language model solutions from different providers. This will enable broader performance comparisons across diverse architectures in ciphertext-only attack scenarios.

The following key direction is exploring more from (classical) cryptography. Studies using diverse types of ciphers will allow comparisons on how certain transformations (both structural and statistical) on the plaintext impact the models' ability to infer meaningful outputs.

Additional extensions may involve testing with longer ciphertext inputs, varying the number of unique guesses produced, and improving prompt construction strategy. For example, the generalizability of fine-tuning a language model on encrypted data or text formatted like a cipher may yield novel results in developing better decryption heuristics with minimal information.

VI. Conclusion

This study explores a ciphertext-only attack approach against classical ciphers, leveraging Large Language Models (LLMs). Our findings indicate a notable inconsistency in the LLM agents' performance. However, despite initial failures, repeated attempts yielded successful attacks. This inconsistency raises significant concerns regarding the potential security implications. While cryptanalysis encompasses a broader spectrum of techniques, our results suggest that LLMs are emerging as a relevant tool in this domain. We posit that it is only a matter of time before LLMs can effectively execute more sophisticated attacks, such as linear and differential cryptanalysis, which are well-established in the cryptography community.

Acknowledgment

This work was partly supported by the Faculty of Computer Science and Engineering at Ss. Cyril and Methodius University in Skopje, and by the European Union's ERASMUS MUNDUS program under the Cyber-MACS project (Project #101082683, https://cybermacs.eu/).

References

[1] N. Fürthauer, V. Mikhalev, N. Kopal, B. Esslinger, H. Lampesberger, and E. Hermann, "Evaluating Deep Learning Techniques for Known-Plaintext Attacks on the Complete Columnar Transposition Cipher," *International Conference on Historical Cryptology*, pp. 82–90, Jun. 2022. [Online]. Available: https://ecp.ep.liu.se/index.php/histocrypt/article/view/394

[2] M. Stamp, "A Revealing Introduction to Hidden Markov Models," in *Introduction to Machine Learning with Applications in Information Security*, 1st ed. Chapman and Hall/CRC, Sep. 2017, pp. 7–35.

[3] R. Vobbilisetty, F. Di Troia, R. M. Low, C. A. Visaggio, and M. Stamp, "Classic cryptanalysis using hidden markov models," *Cryptologia*, vol. 41, no. 1, pp. 1–28, 2017.
[4] D. Charan, "Generative Adversarial Networks for Classic Cryptanalysis," *Masters Projects*, Sep. 2021. [Online]. Available: https://scholarworks.sjsu.edu/etd_projects/1034
[5] A. N. Gomez, S. Huang, I. Zhang, B. M. Li, M. Osama, and Kaiser, "Unsupervised Cipher Cracking Using Discrete," *ICLR*, 2018.
[6] X. Zhuang and A. Yan, "Deep-learning-based ciphertext-only attack on optical scanning cryptosystem," *Optics & Laser Technology*, vol. 157, p. 108744, 2023.
[7] E. Firouzi, M. Ghafari, and M. Ebrahimi, "Chatgpt's potential in cryptography misuse detection: A comparative analysis with static analysis tools," in *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 2024, pp. 582–588.
[8] Z. Masood and M. V. Martin, "Beyond static tools: Evaluating large language models for cryptographic misuse detection," *arXiv preprint arXiv:2411.09772*, 2024.
[9] N. Sugio, "Implementation of cryptanalytic program for ascon using chatgpt," in *2024 Twelfth International Symposium on Computing and Networking Workshops (CANDARW)*, 2024, pp. 307–313.
[10] S. Nobuyuki, "Implementation of cryptanalytic programs using chatgpt," *IACR Cryptol. ePrint Arch.*, 2024.
[11] H. Kwon, M. Sim, G. Song, M. Lee, and H. Seo, "Novel approach to cryptography implementation using chatgpt," *IACR Cryptol. ePrint Arch.*, vol. 2023, p. 606, 2023.

Implementation of Non-Interactive Zero-Knowledge Proof Based Authentication Protocol for Student Authentication

Zivko Atanaskoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
zivko.atanaskoski@finki.ukim.mk

Zorica Karapancheva
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
zorica.karapancheva@finki.ukim.mk

Vesna Dimitrova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
vesna.dimitrova@finki.ukim.mk

Riste Stojanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
riste.stojanov@finki.ukim.mk

***Abstract*—In this paper, we present a privacy-preserving authentication protocol based on non-interactive zero-knowledge proofs (NIZKs) tailored for student identity verification. The proposed system enables students to be validated by other parties without revealing their student ID, using zk-SNARKs constructed with Circom and verified via SnarkJS. The protocol leverages the Poseidon hash function to efficiently encode private inputs within zero-knowledge circuits. A trusted setup is used to generate proving and verification keys based on the Groth16 scheme. The implementation includes three phases such as input preparation, witness and proof generation, and proof verification against a public list of valid students. The verifier, without knowledge of the private data, can confirm the validity of a student's request, making the system suitable for secure integration in third-party services such as transportation or educational platforms. Our work demonstrates that NIZKs can offer efficient, scalable, and privacy-respecting authentication solutions in real-world use cases.**

***Index Terms*—Zero-Knowledge Proofs (ZKPs), Non-Interactive Zero-Knowledge (NIZK), zk-SNARK, Circom, SnarkJS, Poseidon Hash, Privacy-Preserving Authentication, Structured Reference String (SRS), Student Verification, Groth16.**

I. Introduction

The rapid advancement of digital services demands secure and privacy-preserving authentication methods. Traditional verification approaches often require the disclosure of sensitive personal information, which can lead to significant security risks [1]. Zero-knowledge proofs (ZKPs) provide a compelling alternative by enabling a user to prove the truth of a statement without revealing any underlying secret information [2]. However, the interactive nature of conventional ZKPs introduces complexity and inefficiencies in practical implementations.

To address these challenges, this work applies an established non-interactive zero-knowledge proof (NIZK) protocol that leverages a Common Reference String (CRS) model to achieve efficient, single-step proof generation and verification [3]. Our approach utilizes proven cryptographic primitives—most notably, the Poseidon hash function—to construct a system where a prover can generate a proof that validates their claim without disclosing sensitive details [4]. The implementation is built on a modern toolchain comprising Node.js, Circom, and SnarkJS, reflecting contemporary best practices in zero-knowledge proof system design.

By integrating these robust techniques, our implementation demonstrates that privacy-preserving authentication can be achieved efficiently, enabling secure applications across various domains.

The rest of this paper is organized as follows: Related work is reviewed in Section II. In Section III, we introduce the key concepts of non-interactive zero-knowledge proofs. The system design and architecture are detailed in Section IV, followed by the implementation aspects in Section V. Limitations are discussed in Section VI, and future work is outlined in Section VII. Finally, Section VIII concludes the paper.

II. Related work

Zero-knowledge proofs (ZKPs) were developed by Goldwasser, Micali, and Rackoff as a fundamental cryptographic concept [1], while Goldreich, Micali, and Wigderson [5] demonstrated that ZKPs apply to all nondeterministic polynomial time statements (NP). Initially designed for two-party interactions, these proofs are essential in multiparty cryptographic protocols, specifically in evaluating secure functionss [6]. However, interactive ZKPs have downsides, including flexibility, which allows an adversary to create new proofs by manipulating current ones, which requires a greater communication cost [7]. Concurrent execution also introduces risks, as multiple verifiers can collaborate to extract information, requiring polylogarithmic rounds to maintain security [8],

[9]. To address these challenges, Blum, Feldman and Micali [10] introduced non-interactive zero-knowledge proofs (NIZK) in the common random string model, demonstrating how to construct computational NIZK proof systems to prove a single statement about any NP language. The first computational NIZK proof system for multiple theorems was later developed by Blum, De Santis, Micali, and Persiano [11], depending on the number-theoretic assumptions, such as the complexity of deciding quadratic residues modulo a composite number. Feige, Lapidot, and Shamir [12] extended this research by constructing NIZK proofs based on any trapdoor permutation. Despite NIZKs advantages, ensuring non-malleability remained a challenge. To address this, Sahai [13] introduced non-malleable NIZK, initially restricted to a limited number of proofs and later extended to an unlimited environment using trapdoor permutations. Much research has been dedicated to the construction of efficient NIZK proofs [14]–[16], with the hidden random bits method being the most popular approach. NIZK eliminates the interactive ZKP issue of concurrent execution, although flexibility remains an issue and demands cautious protocol design [7], [17]–[19]. Additionally, NIZK proofs have proven crucial in encryption schemes resistant to adaptive chosen-ciphertext attacks (CCA2), improving cryptographic security [20]. In order to address the increasing needs of cryptographic applications, both interactive and non-interactive ZKPs are constantly evolving and achieving balance between performance and security.

III. Non-Interactive Zero Knowledge Proof

Zero-Knowledge Proofs (ZKPs) are cryptographic methods that allow one party, the prover, to convince another party, the verifier, that they know a certain piece of information without revealing any details about the information itself. This feature makes ZKPs particularly useful in privacy-preserving applications such as authentication, secure computations, and blockchain technologies [21]. There are both advantages and disadvantages to this protocol compared to other authentication schemes, as well as challenges that must be addressed to make it practical for general use [22]. A fundamental aspect of ZKPs is that they must achieve three key properties [23]:

1) Completeness – If the statement is true, an honest verifier will be convinced by an honest prover.
2) Soundness – If the statement is false, no dishonest prover can convince an honest verifier except with a insignificant probability.
3) Zero-Knowledge – The verifier learns nothing about the secret information beyond its validity.

Traditional zero-knowledge proofs (ZKPs) were initially designed as interactive protocols, requiring multiple rounds of communication between the prover and the verifier. While interactive ZKPs offer strong security guarantees, they are often impractical for large-scale, decentralized, or asynchronous systems [21]. Non-Interactive Zero-Knowledge Proofs (NIZKs) overcome this limitation by eliminating the need for repeated interaction [24]. By relying on a Common Reference String (CRS), NIZKs allow a prover to generate a single proof that can be independently verified by multiple verifiers at any time. This makes NIZKs highly scalable, efficient, and better suited for real-world applications such as secure authentication and blockchain privacy mechanisms. In contrast to the interactive version, NIZKs reduce computational overhead and reduce the risks associated with network delays or interference during the proof exchange. NIZKs are built upon several core cryptographic principles:

1) Common Reference String (CRS) – A predefined string shared between the prover and verifier that facilitates proof generation and verification.
2) Witness – The secret knowledge the prover possesses and wishes to prove knowledge of.
3) Statement – A public assertion that the prover wants to convince the verifier about.
4) Proof – A cryptographic string generated by the prover that convinces the verifier of the validity of the statement.
5) Verification Algorithm – A function that checks the validity of the proof without revealing additional information about the witness.

IV. System Design and Architecture

For the purpose of implementing a NIZK use case, we designed an application following the NIZK concepts focusing on verifying the identity of a student without exposing any sensitive information such as their student ID. The design enables the student to prove their identity by generating a cryptographic proof that validates their claim, without revealing the actual private data to the verifier. This approach ensures privacy and security while maintaining integrity in the verification process. The application uses cryptographic primitives that are optimized for zk-SNARK (Zero-Knowledge Succinct Non-Interactive Arguments of Knowledge) circuits, providing an efficient and secure method of identity verification [25].

The authentication process follows a structured protocol as shown on Figure 1:

1) **Input Preparation** - The student provides the private data input (their student ID) which then is hashed and turned into a public representation. The hash value is used as a public signal that will later be compared with the verifiers database.

   In the phase of enrollment of a student in a faculty, the faculty generates a student ID unique to each student. In parallel, the new student ID as a hashed version is added to a list (database) of valid student hash values available to third-party organizations available for public use.

2) **Witness and Proof Generation** - Intermediate values, known as witnesses, are computed to ensure that the private data correctly corresponds to the public hash. Then a proof is generated from the public representation, which confirms the validity of the student's request for authentication.

When the student needs to authenticate themselves as valid students to organizations outside the faculty without providing their student ID, they need to provide a Proof that would validate them. For that purpose, the request of the authentication is sent to a Prover which calculates the Proof.

3) **Verification** - The verifier checks the authenticity of the proof by comparing it to public signals, such as a hash of the data, ensuring that the student's request is legitimate without ever having access to the underlying private data (their student ID).

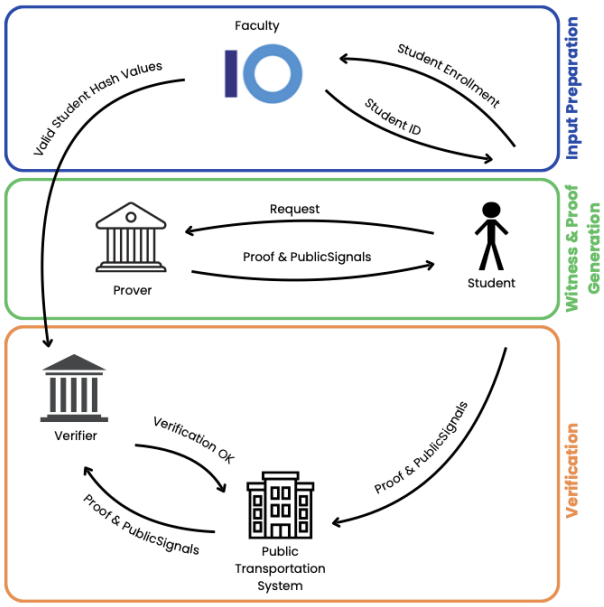


Fig. 1. Use-case-level student authentication process

In terms of the student trying to authenticate himself to an organization (such as the Public Transportation System), the student needs to send the Proof generated in the previous step together with the hashed version of the ID known as Public Signal. The organization then forwards the proof to a verifier which has the list of valid student hash values and has the capability to verify the Proof. As the last step, the verifier confirms or denies the authentication of the student. In this step, the verifier nor the organization have the private data of the student but using NIZK manages to authenticate the student.

V. Implementation Details

To implement a privacy-preserving student identification system, we developed a Node.js application that utilizes non-interactive zero-knowledge proofs (NIZKs) based on the Groth16 proving system [26]. This approach enables the system to verify that a user possesses a valid student ID without revealing the actual identifier. The implementation relies on Circom for circuit construction, SnarkJS for the cryptographic workflow, and the Poseidon hash function for privacy-preserving input encoding [27]. A visual representation of the workflow summary can be seen in Figure 2. The complete code implementation is publicly accessible at the repository cited in [28].

Preliminary Setup: Circuit Initialization and Trusted Setup

Circuit and Hash Function

The core logic is defined in a Circom circuit, which takes a private student ID as input and computes a public Poseidon hash output. The Poseidon hash was chosen for its efficiency and compatibility with arithmetic circuits. To ensure consistency between the frontend and the circuit, we use the `circomlibjs` library to compute the same Poseidon hash during input preparation [4], [27].

Trusted Setup: Groth16

The Groth16 proving system requires a trusted setup consisting of two phases. In the first phase, known as the Powers of Tau ceremony, we generate a structured reference string (SRS) that is circuit-independent. In the second phase, we use this SRS to produce proving and verification keys tailored to our specific circuit [26]. These keys are stored in `.zkey` and `.json` files respectively. Additionally, the circuit is compiled into a Rank-1 Constraint System (`.r1cs`) and a WebAssembly binary (`.wasm`) for witness generation.

Phase 1: Input Preparation

Once the student ID is provided by the user, the following steps are performed:

- The student ID is processed through the Poseidon hash function to generate a public representation.
- The computed hash, along with the original input, is stored in a structured input file (`input.json`). This file serves as the basis for the subsequent witness generation.

Phase 2: Witness and Proof Generation

Once the student ID is provided and its Poseidon hash is stored in the input file (`input.json`), the compiled circuit processes this file to generate a witness file (`witness.wtns`) that contains the intermediate values satisfying the circuit constraints. This witness, together with the proving key, is then used to generate a zero-knowledge proof along with the corresponding public signals—which include the Poseidon hash. The resulting proof and public outputs are stored in `proof.json` and `public.json`, respectively, and are transmitted to the verifier.

Phase 3: Verification

On the verifier side, the Poseidon hash received as a public signal is checked against a public list of valid student hashes stored in `trusted_hashes.json`. This list does not reveal any personal information. Once a match is found, the Groth16 verification algorithm is executed using the verification key,

proof, and public signals. If the proof is valid, the user is authenticated without any disclosure of their original ID.

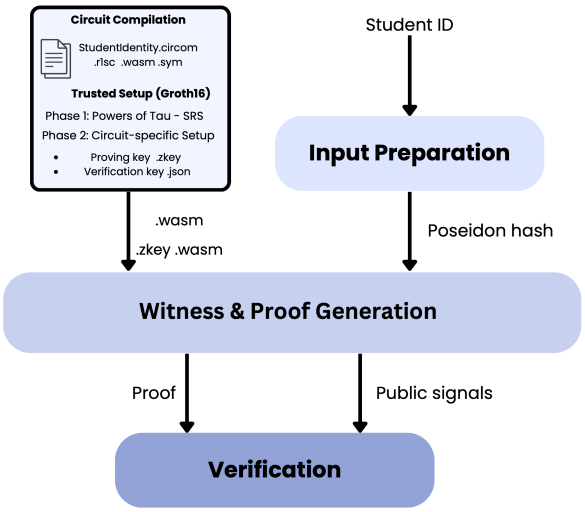


Fig. 2. Implementation-level workflow of the NIZK protocol

Workflow Summary

1) **Input Preparation** – The student ID is hashed using the Poseidon function.
2) **Witness & Proof Generation** – A witness is generated from the circuit and input. Then a zk-SNARK proof and public output are created.
3) **Verification** – The proof is verified and the public hash is validated against a trusted list.

VI. Limitations

While the proposed NIZK-based student authentication protocol offers strong privacy guarantees and efficient non-interactive verification, several limitations and potential challenges remain [29], [30]:

- **Trusted Setup Requirement:** The Groth16 proving system requires a trusted setup phase. If the randomness (commonly referred to as ”toxic waste”) is not properly discarded or securely generated, a malicious actor could potentially generate fake proofs. This requirement introduces a dependency on the integrity of the setup ceremony.
- **Scalability in Large Deployments:** In large-scale systems with thousands of users, maintaining and synchronizing the list of valid Poseidon hashes can become challenging. Efficient set membership proofs or Merkle tree-based alternatives may be required for more scalable deployments.
- **Proof Generation Overhead:** Although proof verification is efficient, generating zk-SNARK proofs can be computationally intensive, especially on mobile or edge devices. This may limit usability in lightweight environments without optimization or hardware acceleration.
- **Limited Revocation Capabilities:** The system currently lacks a robust mechanism to revoke access or invalidate existing proofs after student status changes. Without dynamic proof schemes or expiration strategies, managing revocation becomes manual and error-prone.
- **Circuit Rigidity:** Proofs are tied to a specific arithmetic circuit and proving key. Any logic changes, such as modifying constraints or adding features, require recompiling the circuit and performing a new trusted setup, which can be operationally complex.
- **Replay Attack Risk:** In its current form, the system does not include nonce-based or session-based protections. This could allow a previously valid proof to be reused (replayed) by an unauthorized party unless contextual binding (e.g., timestamp or challenge-response) is introduced.

VII. Future Work

To improve scalability and security, future versions of the system may adopt transparent zk-SNARKs such as PLONK to remove the need for a trusted setup. Integrating Merkle trees for student list commitment would allow efficient set membership proofs and better scalability. Optimizations for proof generation on low-power devices could enhance usability in mobile environments. To address replay attacks, context-bound proofs using nonces or timestamps should be introduced. Additionally, incorporating expiration or revocation mechanisms into proofs would ensure better control over authentication validity in dynamic real-world scenarios.

VIII. Conclusion

This paper presents a privacy-preserving authentication system using non-interactive zero-knowledge proofs tailored for student identity verification. By leveraging zk-SNARKs, the Poseidon hash function, and the Groth16 proving system, the proposed solution allows secure authentication without disclosing private data. The implementation demonstrates that modern ZK toolchains like Circom and SnarkJS can be effectively used to build scalable and secure applications. Future enhancements may include transitioning to transparent SNARKs, integrating Merkle tree-based verification, and introducing revocation and replay protection mechanisms. The approach has potential for broader applications in privacy-first authentication systems across various sectors.

ACKNOWLEDGMETS

This work was partly supported by the Faculty of Computer Science and Engineering at Ss. Cyril and Methodius University in Skopje.

REFERENCES

[1] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexity of interactive proof-systems," in *Providing sound foundations for cryptography: On the work of shafi goldwasser and silvio micali*, 2019, pp. 203–225.

[2] O. Goldreich, *Foundations of Cryptography, Volume 2*. Cambridge university press Cambridge, 2004.

[3] A. De Santis, S. Micali, and G. Persiano, "Non-interactive zero-knowledge proof systems," in *Advances in Cryptology—CRYPTO'87: Proceedings 7*. Springer, 1988, pp. 52–72.

[4] L. Grassi, D. Khovratovich, C. Rechberger, A. Roy, and M. Schofnegger, "Poseidon: A new hash function for {Zero-Knowledge} proof systems," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 519–535.

[5] O. Goldreich, S. Micali, and A. Wigderson, "Proofs that yield nothing but their validity or all languages in np have zero-knowledge proof systems," *Journal of the ACM (JACM)*, vol. 38, no. 3, pp. 690–728, 1991.

[6] ——, "How to play any mental game, or a completeness theorem for protocols with honest majority," in *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, 2019, pp. 307–328.

[7] D. Dolev, C. Dwork, and M. Naor, "Non-malleable cryptography," in *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, 1991, pp. 542–552.

[8] R. Canetti, J. Kilian, E. Petrank, and A. Rosen, "Black-box concurrent zero-knowledge requires\tilde {Ω}(log n) rounds," in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 2001, pp. 570–579.

[9] J. Kilian and E. Petrank, "Concurrent and resettable zero-knowledge in poly-loalgorithm rounds," in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 2001, pp. 560–569.

[10] M. Blum, P. Feldman, and S. Micali, "Non-interactive zero-knowledge and its applications," in *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, 2019, pp. 329–349.

[11] M. Blum, A. De Santis, S. Micali, and G. Persiano, "Noninteractive zero-knowledge," *SIAM Journal on Computing*, vol. 20, no. 6, pp. 1084–1118, 1991.

[12] U. Feige, D. Lapidot, and A. Shamir, "Multiple noninteractive zero knowledge proofs under general assumptions," *SIAM Journal on computing*, vol. 29, no. 1, pp. 1–28, 1999.

[13] A. Sahai, "Non-malleable non-interactive zero knowledge and adaptive chosen-ciphertext security," in *40th annual symposium on foundations of computer science (Cat. No. 99CB37039)*. IEEE, 1999, pp. 543–553.

[14] I. Damgård, "Non-interactive circuit based proofs and non-interactive perfect zero-knowledge with preprocessing," in *Workshop on the Theory and Application of of Cryptographic Techniques*. Springer, 1992, pp. 341–355.

[15] J. Kilian and E. Petrank, "An efficient noninteractive zero-knowledge proof system for np with general assumptions," *Journal of Cryptology*, vol. 11, pp. 1–27, 1998.

[16] J. Boyar, I. Damgård, and R. Peralta, "Short non-interactive cryptographic proofs," *Journal of Cryptology*, vol. 13, pp. 449–472, 2000.

[17] M. Naor and M. Yung, "Public-key cryptosystems provably secure against chosen ciphertext attacks," in *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, 1990, pp. 427–437.

[18] M. Bellare and S. Goldwasser, "New paradigms for digital signatures and message authentication based on non-interactive zero knowledge proofs," in *Conference on the Theory and Application of Cryptology*. Springer, 1989, pp. 194–211.

[19] S. Goldwasser and R. Ostrovsky, "Invariant signatures and non-interactive zero-knowledge proofs are equivalent," in *Annual International Cryptology Conference*. Springer, 1992, pp. 228–245.

[20] A. De Santis, G. Di Crescenzo, and G. Persiano, "Necessary and sufficient assumptions for non-interactive zero-knowledge proofs of knowledge for all np relations," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2000, pp. 451–462.

[21] A. De Santis, G. Di Crescenzo, R. Ostrovsky, G. Persiano, and A. Sahai, "Robust non-interactive zero knowledge," in *Advances in Cryptology—CRYPTO 2001: 21st Annual International Cryptology Conference, Santa Barbara, California, USA, August 19–23, 2001 Proceedings 21*. Springer, 2001, pp. 566–598.

[22] R. S. Norville, K. Namuduri, and R. Pendse, "Node authentication in networks using zero-knowledge proofs," in *Web Services Security and E-Business*. IGI Global, 2007, pp. 142–164.

[23] X. Tang, L. Shi, X. Wang, K. Charbonnet, S. Tang, and S. Sun, "Zero-knowledge proof vulnerability analysis and security auditing," *Cryptology ePrint Archive*, 2024.

[24] J. Groth, R. Ostrovsky, and A. Sahai, "New techniques for noninteractive zero-knowledge," *Journal of the ACM (JACM)*, vol. 59, no. 3, pp. 1–35, 2012.

[25] M. Petkus, "Why and how zk-snark works," *arXiv preprint arXiv:1906.07221*, 2019.

[26] H. Lipmaa, "Polymath: Groth16 is not the limit," in *Annual International Cryptology Conference*. Springer, 2024, pp. 170–206.

[27] M. Bellés-Muñoz, M. Isabel, J. L. Muñoz-Tapia, A. Rubio, and J. Baylina, "Circom: A circuit description language for building zero-knowledge applications," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 6, pp. 4733–4751, 2022.

[28] https://github.com/atanaskoskizivko/ZKP, accessed: 1-Apr-2025.

[29] E. Ben-Sasson, A. Chiesa, D. Genkin, E. Tromer, and M. Virza, "Snarks for c: Verifying program executions succinctly and in zero knowledge," in *Annual cryptology conference*. Springer, 2013, pp. 90–108.

[30] E. Ben-Sasson, A. Chiesa, E. Tromer, and M. Virza, "Scalable zero knowledge via cycles of elliptic curves," *Algorithmica*, vol. 79, pp. 1102–1160, 2017.

Onilne Shoppers Purchasing Intention Prediction and Analysis Using Machine Learning

Natalija Arizankoska
FEEIT
Ss. Cyril and Methodius University in Skopje, North Macedonia
natalijaarizankoska@gmail.com

Hristijan Gjoreski
FEEIT
Ss. Cyril and Methodius University in Skopje, North Macedonia
hristijang@feit.ukim.edu.mk

***Abstract*— Online shopping has rapidly grown in popularity, offering consumers the convenience of purchasing products from the comfort of their homes. With the rise in e-commerce, especially for groceries, the ability to avoid long lines and save time has made it a preferred choice for many. In this paper, we explore the use of customer features, primarily based on their interests and resources, to predict whether they will complete a purchase at the end of their browsing session. Machine Learning classification was used to make these predictions, and the performance was evaluated using accuracy, precision, recall, and F1 score, with a particular focus on the analysis of the Precision-Recall curve which shows how these models handle imbalanced datasets. The results show that Random Forest achieved the highest accuracy (0.9007), precision (0.7433), and F1 score (0.6458). Other models, including SVC, XGBoost, and AdaBoost, produced comparable but slightly worse performance metrics. Additionally, clustering models such as K-Means and DBSCAN were used to identify patterns and outliers, revealing important insights into customer behavior and segmentation. These findings underscore the importance of integrating advanced analytics into customer segmentation strategies, ultimately enabling more targeted and effective online marketing efforts.**

Keywords*— *Online shopping, e-commerce, purchase prediction, classification models, clustering models, imbalanced datasets

I. Introduction

Online shopping has transitioned from a specialized service into an everyday norm, with consumers increasingly preferring e-commerce over physical retail stores [1]. Its appeal lies in the ability to browse a broad global catalogue 24/7, avoid crowds and queues and rely on detailed reviews when making purchase decisions, all from the comfort of their home. Despite these advantages, online retail faces challenges such as fraud, damaged or missing deliveries, and the inability to inspect products physically.

In a typical e-commerce journey, users browse items, compare specifications and reviews, add products to a virtual cart, and complete secure, web-based payments before delivery. Regardless of the seamless experience for the buyers, roughly 70% of the shopping carts are abandoned resulting in nearly 4 trillion USD being lost in annual sales [2],[3]. Because only a fraction of visitors ever reach checkout, broad discounting offers often waste marketing budget, where personalized promotions for high-intent shoppers can deliver a noticeable higher return of investment [3].

Applying machine learning algorithms for predicting purchases on real-time click stream data can help retailers to recapture value. Some of the proposed methods for purchase intention prediction presented in this paper are the following:

- Decision Trees which recursively split sessions based on feature thresholds (e.g., time spent on certain pages, types of pages browsed), creating a hierarchy of purchase vs. nonpurchase decisions.
- Random Forests aggregate such trees trained on bootstrap samples, boosting robustness.
- Support Vector Classifiers (SVC) project sessions into a high-dimensional space and find the hyperplane that best separates purchasers from non-purchasers.
- AdaBoost sequentially fits weak learners, reweighing misclassified sessions.
- XGBoost uses gradient-based boosting with regularization and optimized tree construction, delivering high accuracy and scalability.

Besides classification, unsupervised clustering methods will be introduced later in this paper revealing browsing patterns:

- K-Means partitions session into k clusters by minimizing within cluster variance, helping identify distinct shopper segments.
- DBSCAN groups sessions that are based on density, detecting randomly shaped clusters and isolating noise.

These techniques can be used to accurately forecast purchases and enable businesses to launch individualized campaigns, optimize inventory, and prioritize loyalty when it comes to frequency buyers and thereby boost conversion rates customer satisfaction and ROI (Return Of Investment).

II. Related Work

In online shopping purchase-intent prediction, Baati and Mohsil introduce a real-time system that leverages only session and visitor attributes such as aggregated page views and user demographics to train Naïve Bayes, C4.5 Decision Tree and with SMOTE oversampling, a Random Forest that achieves an accuracy of 86.78% and a F1 score of 60% [4].

Expanding on feature representation, Kiki and Houndji engineered a rich set of session-level attributes such as total pages viewed, average durations per visit and binary indicators for weekends and month transitions. Using these features, they demonstrated their gradient boosting classifier, which surpasses neural network baselines in both precision and F1 score [5]. In their most recent study, Abdullah-All-Tanvir and his co-authors integrate χ^2 based feature selection using SMOTE (Synthetic Minority Over-Sampling Technique) technique to isolate the most informative session attributes, enabling their XGBoost model to achieve 90.65% accuracy, a 0.937 AUROC (Area Under Receiver-Operating Characteristic Curve), and a 0.754 AUPRC (area under precision-recall curve) demonstrating the impact of targeted feature reduction when paired with boosting techniques for online shopping intent prediction [6].

III. Dataset

The dataset used in this project is publicly available on UC Irvine Machine Learning Repository consisting of 12, 330 samples [7]. Each sample is formed by sessions that belong to different users in a 1-year period just to avoid cases like campaigns, personal special days or periods. The dataset itself consists of 18 features, 10 of which are numerical and 8 of them are categorical. The 'Revenue' label is used as the target feature to predict whether a purchase was made or not. Importantly, this is a very imbalanced dataset with approximately 84% of the sessions have Revenue=False, while only 16% have Revenue=True. Below, we enumerate all 18 features as a bulleted list, providing concise explanation for each:

- Administrative, Informational, Product Related: The number of pages that the shopper visited in a session related to administrative (informal, product related) information.
- Administrative, Informational, Product Related Duration: Time spent by the shopper on these kind of pages (in seconds).
- Bounce Rate: Average bounce rate of the pages browsed by the shopper. (Bouncing – when a user enters a site and leaves without triggering any other requests).
- Exit Rate: Average exit rate value of the pages that are browsed by the shopper.
- Page Value: The average number of web pages that the user visited before completing an e-commerce transaction.
- Special Day: How close is the site visiting time to a specific special day (Mother's Day, Valentine's Day).
- Operating System: The operating system used by the current site visitor.
- Browser: Internet Browser the online shopper is using.
- Location: Geographic location from which the browsing session has been started by the online shopper.
- Traffic Type: Traffic source by which the online shopper has reached the shopping site.
- Visitor Type: Returning, New or other category.
- Weekend: Binary value representing whether the day of the browsing is weekend.
- Month: Month value of the browsing date.
- Revenue: Class label representing whether the visit has been finalized with a purchase or not.

IV. Data Preprocessing

In this study, data preprocessing and exploratory data analysis (EDA) were performed to ensure the dataset was clean, structured, and ready for classification modeling.

The first step towards a clean dataset was checking for duplicate entries to eliminate redundant records that could bias the model. To further enhance data quality, outlier analysis was performed to detect and handle extreme values that could distort model predictions. This involved visualizing distributions using box plots shown below on Fig. 1. Standardization of data was then carried out to bring all features to a uniform scale, particularly for numerical attributes with varying ranges.

Additionally, skewness (a degree of asymmetry observed in a probability distribution) [8] correction was applied to ensure a balanced distribution of numerical variables.

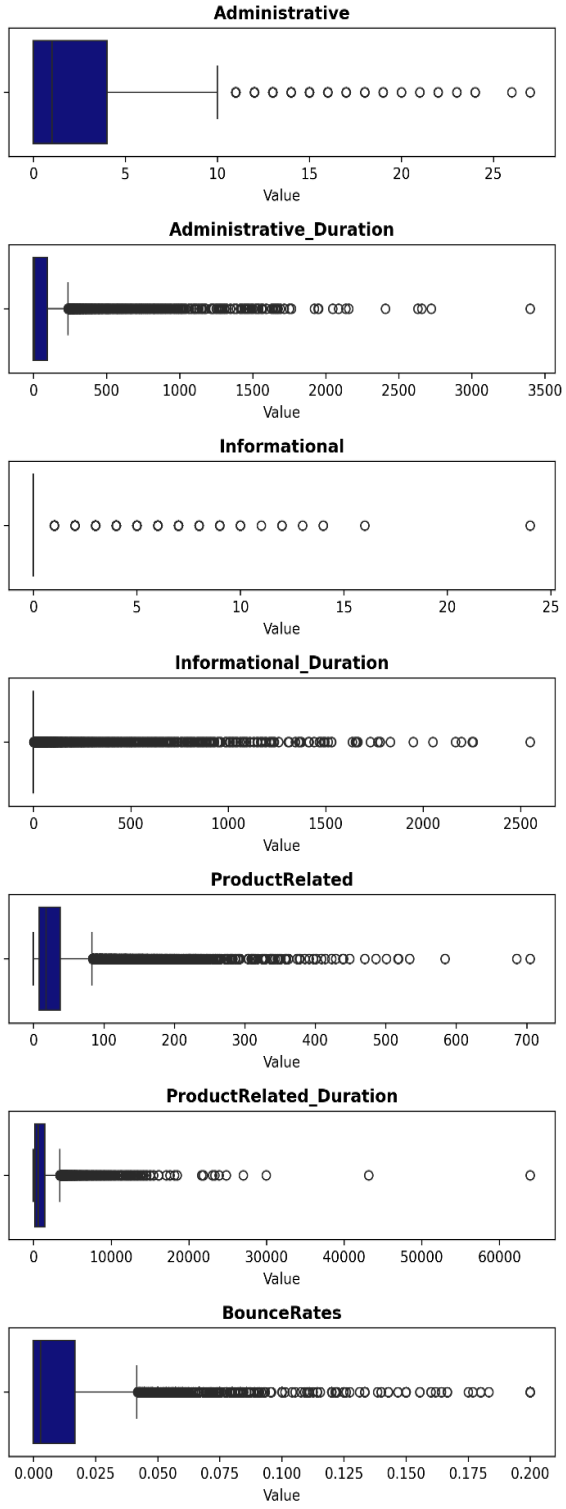


Fig. 1: A boxplot illustrating the distribution of the data for various features, highlighting the presence of potential outliers.

Since highly skewed data can negatively impact model performance, transformation techniques such as logarithmic transformations were utilized to bring the skewness within an acceptable range.

Another critical aspect of data preprocessing was handling correlation between features. Highly correlated features were identified using correlation matrices, and redundant variables were dropped to prevent model overfitting and redundancy in feature importance.

Through extensive data preprocessing and exploratory data analysis (EDA) including plotting several features against the target variable 'Revenue' (Fig. 2, 3), the dataset was refined to enhance the predictive power of the classification algorithms. These steps ensured that the models could learn meaningful patterns from the data without being affected by irrelevant features.

Furthermore, the dataset was found to be imbalanced, with the false class (non purchasing customers) significantly outnumbering the true class (Fig. 4). This imbalance posed a challenge for classification models, as they tend to favor the dominant class, leading to biased predictions. To avoid this issue, feature selection was performed to identify the most relevant attributes. Out of the 17 features, 13 were chosen by computing the ANOVA F-value and resulting with me groups of features that had the most valuable information. By refining the dataset through careful selection and balancing techniques, the classification models were better equipped to recognize patterns in purchasing behavior and improve predictive accuracy.

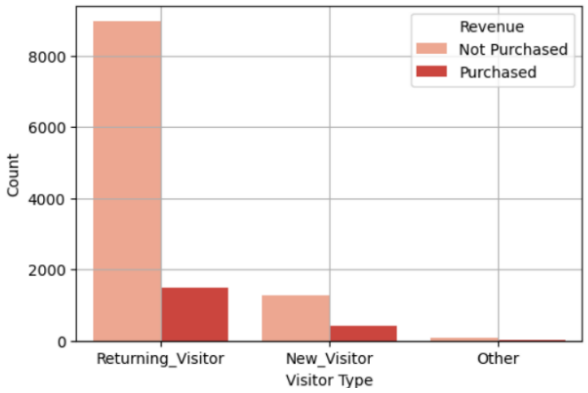


Fig. 2: Revenue generated by different types of visitor categories.

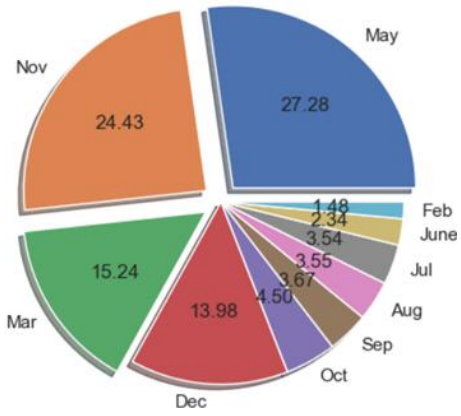


Fig. 3: A visualization of monthly revenue distribution, illustrating how revenue fluctuates across different months.

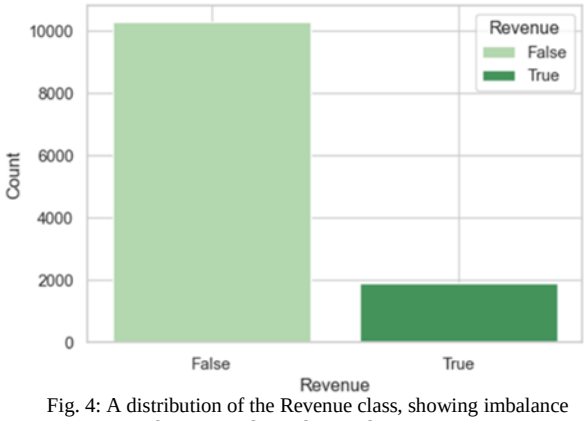


Fig. 4: A distribution of the Revenue class, showing imbalance between False and True classes.

V. Methods

The feature vectors were used as input to classification algorithms for them to learn to correctly classify whether a purchase was made or not. Besides the classification models, two of the clustering algorithms were used as well. In this project, 5 machine learning algorithms were employed to predict online shopping purchase intentions. The algorithms applied were: Decision Tree, Support Vector Classifier (SVC), Random Forest, XGBoost, and AdaBoost.

The Decision Tree model classifies the data based on simple decision rules resulting from data features and by that creating a tree like structure [9], while the Support Vector Classifier finds the hyperplane that best separates classes by maximizing the margin between them [10]. Random Forest, an ensemble of Decision Trees, aggregates predictions from multiple trees to enhance accuracy [11]. XGBoost (Extreme Gradient Boosting), an optimized implementation of gradient boosting, builds an ensemble of decision trees sequentially and with each tree it attempts to correct the mistakes made by the previous one [12], and AdaBoost enhances the accuracy of weak classifiers by iteratively adjusting the weights of misclassified instances [13]. These models were chosen for their ability to handle complex, high-dimensional data and their effectiveness in classification tasks. The underlying mathematical principles of these algorithms focus on minimizing classification error through various techniques like splitting nodes, finding optimal hyperplanes, and combining weak learners to produce stronger predictions.

Cross-validation was applied to the data so that the performance of the classification models would be more reliable. This technique helps to ensure that the models generalize well unseen data by partitioning the data into multiple subsets and evaluating performance on each subset. After applying cross-validation, a slight improvement in accuracy was observed in almost all the classifiers [14]. This indicates that the models benefited from the validation process, which helped in tuning the hyperparameters and reducing overfitting. Overall, cross-validation provided a more robust estimate of model performance, contributing to better generalization.

In addition to exploring classification models, we first applied Principal Component Analysis (PCA) [15] to reduce the dimensionality of the dataset by capturing the most significant variance. This transformation not only condensed the high-dimensional data into a more manageable, lower-dimensional space but also enhanced the performance of our clustering algorithms by mitigating noise and redundancy.

Following the PCA transformation, K-means clustering [16] was implemented to group similar data points by minimizing intra-cluster distances. Likewise, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [17] was applied to the same PCA-reduced dataset to identify dense regions and isolate outliers. Overall, applying PCA before clustering provided a clearer insight into the underlying structure of the data, which was further compared with our classification results.

VI. Evaluation Metrics

In order to asses and compare the predictive performance of our classifiers, we employ both classical point metrics such as accuracy, precision, recall and F1 score and threshold oriented Precision-Recall (PR) curves [18], [19]. While the first group of metrics provides simple decision summaries, PR-curves reveal the full trade-off between precision and recall as the threshold varies, which is a critical advantage when dealing with imbalanced data.

All of the metrics derive from the four outcomes of a binary classifier, which we summarize in the confusion matrix:

- **True Positives (TP):** Number of positive instances correctly classified as positive.
- **True Negatives (TN):** Number of instances correctly classified as negative.
- **False Positives (FP):** Number of negative instances incorrectly classified as positive.
- **False Negatives (FN):** Number of positive instances incorrectly classified as negative.

Using TP, TN, FP and FN we computed:

1. **Accuracy** (1) that measures the overall percentage of correctly classified instances in the dataset.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

2. **Recall** (2) which focuses in the proportion of actual positive instances correctly identified by the model, highlighting its ability to detect true positives.

$$\text{Recall} = TP/(TP+FN) \quad (2)$$

3. **Precision** (3), on the other hand, measures the proportion of predicted positives that are actually correct, emphasizing the model's ability to avoid false positives.

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

4. Finally, the **F1 score** (4) is the harmonic mean between precision and recall, providing a balanced evaluation of the model's performance, particularly in cases where the data is imbalanced.

$$F1 = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

Although important, these metrics can mask important behaviors particularly when the positive class is underrepresented.

A PR-curve plots precision versus recall as the classifier's score from high threshold with high precision and low recall, to liberal with low precision and high recall. The area under the PR-curve summarizes overall performance: larger AUC-PR indicates that on average the model achieves better precision-recall trade-offs across all the thresholds. The reason why PR-curve is chosen over the ROC-curve is because of the heavily imbalanced dataset samples and ROC-curves can be misleadingly optimistic with the low false positive rates on the large negative class mask poor detection of the positive minority. PR-curves focus on the positive class, making them more informative when positives are rare.

VII. Results

As mentioned earlier several classification methods were evaluated: Decision Tree, Random Forest, SVC, XGBoost and AdaBoost. The dataset was imbalanced, with a disproportionate number of instances in the true and false classes, making it crucial to balance performance across multiple metrics such as precision, recall, and F1 score.

- Accuracy: Random Forest achieved the highest accuracy at 90,07%, followed closely by XGBoost at 89.74% and SVC at 89.70%.
- Precision: Random Forest outperformed other classifiers with 74.33% indicating that it made the fewest false positive predictions.
- Recall: Decision Tree exhibited the best recall at 59.10% highlighting its strength in identifying true positives.
- F1 Score: Random Forest also led in F1 Score with 64.58%, reflecting a balanced approach between precision and recall.

While Random Forest consistently outperformed other model, Decision Tree showed the highest recall, indicating its better performance in identifying positive instances (detailed results are presented in Table I below). This suggests that while Random Forest was good avoiding false positives, it missed significant numbers of positives, a trade-off highlighted in the confusion matrix (see Fig. 5).

TABLE I. Comparison of performance metrics

	Decision Tree	SVC	Random Forest	XGBoost	AdaBoost
Accuracy (%)	89.52	89.7	90.7	89.74	89.17
Precision (%)	69.97	70.58	74.33	69.2	67.73
Recall (%)	59.10	57.83	56.93	58.04	58.12
F1 score (%)	61.92	63.8	64.58	64	62.48

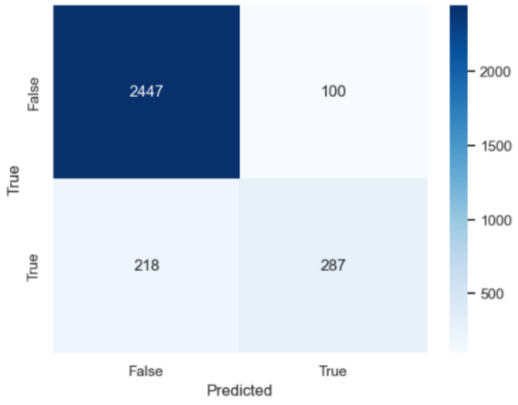


Fig. 5: Confusion Matrix for the Random Forest model.

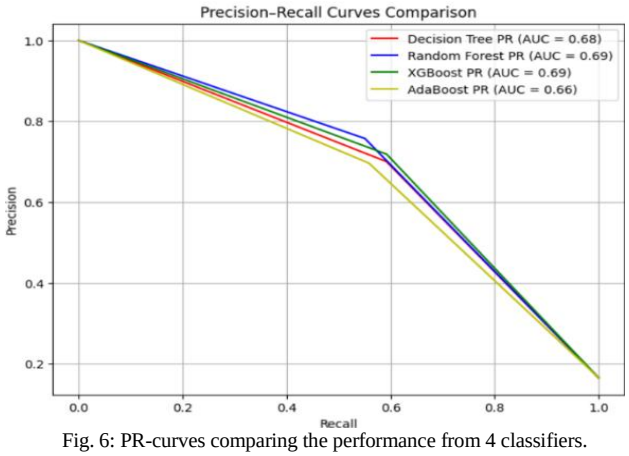


Fig. 6: PR-curves comparing the performance from 4 classifiers.

The PR-curves for the classifiers show their trade-offs between precision and recall. All classifiers showed similar behavior, with Random Forest and XGBoost slightly outperforming Decision Tree and AdaBoost, as indicated by their higher AUC-PR values of 0.69. The PR-curves reveal that Random Forest maintains relatively high precision at moderate recall levels, outperforming other models, especially in high-precision regions. However, as recall increases, precision drops across all models, as expected when trying to identify more positives at the cost of increasing false positives. The curve's shape confirms that Random Forest strikes a good balance between precision and recall, making it the most reliable model for applications where both false positives and false negatives have significant consequences (see Fig. 6).

Following PCA for dimensionality reduction, two clustering algorithms were applied to segment the dataset. K-Means, which assumes spherical clusters, successfully identified 8 well-defined clusters with clear separation (see Fig. 7). This approach is ideal for datasets where clusters are roughly spherical and evenly distributed. However, K-Means struggles when dealing with complex or nonlinear clusters.

On the other hand, DBSCAN, which does not require predefined number of clusters, was able to identify clusters of varying densities and detect noise points to outliers. DBSCAN'S ability to identify irregularly shaped clusters and noise points adds value in scenarios where data may not fit typical cluster shapes.

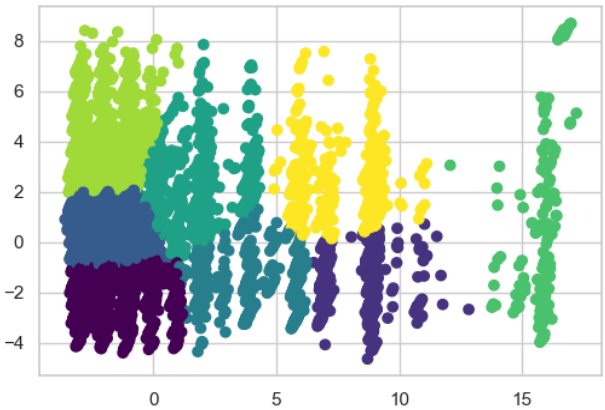


Fig. 7: K-Means cluster representation with the number of clusters set to 8 (K=8).

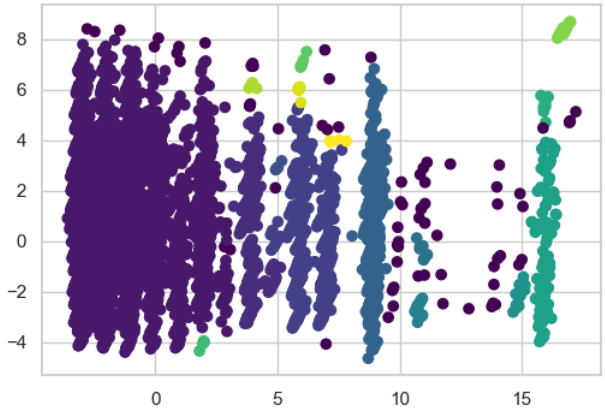


Fig. 8: DBSCAN cluster representation.

The detection of outliers in DBSCAN also suggests that further parameter tuning could improve clustering performance (see Fig. 8). Overall, K-Means provides straightforward solution for well-structured data while DBSCAN excels in identifying noise and handling complex clustering structures.

VIII. Conclusion

This study applied both classification and clustering models after thorough data preprocessing and exploratory data analysis. The classification models exhibited varying trade-offs between precision and recall, emphasizing the challenges of handling imbalanced datasets. Some models excelled in predicting positive cases, while others struggled with recall, highlighting the complexity of achieving a balanced performance in the presence of class imbalances. For clustering, K-Means and DBSCAN were used after PCA for dimensionality reduction. K-Means produced well-defined clusters, effectively segmenting the dataset, while DBSCAN identified clusters of varying densities and highlighted noise points, which is crucial for understanding outliers in imbalanced datasets. The integration of classification and clustering provided complementary insights into the dataset, with classification models assisting in making predictions and clustering uncovering hidden structures and outliers. The findings underscore the importance of selecting the right models based on the data's characteristics, particularly in imbalanced settings, where accuracy, precision, recall and interpretability need to be balanced effectively. Clustering is also critical for uncovering hidden patterns and grouping similar individuals, which is especially valuable in marketing. By segmenting the data into clusters, we can identify people with shared behaviors or interests and create targeted campaigns that appeal to these specific groups. This approach can increase the likelihood of purchases and improve customer engagement by focusing on the right audience with personalized strategies.

To improve model performance in handling imbalanced data, techniques such as SMOTE (Synthetic Minority-Over Sampling Technique) [20] could be applied to augment the minority class and better balance the dataset. SMOTE generates synthetic examples of the minority class, reducing bis toward the majority class and improving recall for the minority class. Additionally, cost-sensitive learning approaches could be explored, where models are penalized more heavily for misclassifying the minority class, aligning the model's training objective with the real-world cost of

misclassification. Furthermore, exploring more advanced algorithms, such as neural networks, could provide significant improvements. Neural networks can capture complex patterns and interactions within the data that basic algorithms might miss, offering better generalization and performance on both precision and recall.

[1] Abdullah-All-Tanvir, Iftakhar Ali Khandokar, A.K.M. Muzahidul Islam, Salekul Islam, Swakkhar Shatabda: A gradient boosting classifier for purchase intention prediction of online shoppers, Heliyon 9 (2023) e15163

[2] V. Rajagopal, G. Malhotra, and S. R. Tiwari, “Utilizing SEM-RFC to predict factors affecting online shopping cart abandonment,” Computers, Materials & Continua, vol. 71, no. 1, pp. 21–35, 2022.

[3] N. Chaudhary and D. R. Chowdhury, “Expanding Click and Buy rates: Exploration of evaluation metrics that measure the impact of personalized recommendation engines on e-commerce platforms,” arXiv preprint arXiv:1901.08901, Jan. 2019.

[4] Baati K, Mohsil M. Real-Time Prediction of Online Shoppers’ Purchasing Intention Using Random Forest. Artificial Intelligence Applications and Innovations. 2020 May 6;583:43–51. doi: 10.1007/978-3-030-49161-1_4. PMCID: PMC7256375.

[5] Y. Kiki and V. R. Houndji, "Prediction of the purchase intention of users on e-commerce platforms using gradient boosting," International Journal of Engineering and Advanced Technology (IJEAT), vol. 10, no. 1, pp. 446-450, Oct. 2020. doi: 10.35940/ijeat.A19291010120.

[6] A.-A. Tanvir, I. A. Khandokar, A. K. M. M. Islam, S. Islam, and S. Shatabda, "A gradient boosting classifier for purchase intention prediction of online shoppers," Heliyon, vol. 9, no. 4, p. e15163, 2023. doi: 10.1016/j.heliyon.2023.e15163.

[7] Sakar, C. & Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5F88Q.

[8] David P. Doane & Lori E. Seward (2011) Measuring Skewness: A Forgotten Statistic?, Journal of Statistics Education, 19:2, DOI: 10.1080/10691898.2011.11889611

[9] Bahzad Taha Jijo, Adnan Mohsin Abdulazeez: Classification Based on Decision Tree Algorithm for Machine Learning, Vol. 02, No. 01, pp. 20 – 28 (2021)

[10] Richard G. Brereton* and Gavin R. Lloyd: Support Vector Machines for classification and regression, Analyst, 2010, 135, 230–267

[11] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25.2 (2016): 197-227.

[12] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[13] Schapire, Robert E. "Explaining adaboost." In Empirical inference: festschrift in honor of vladimir N. Vapnik, pp. 37-52. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[14] Berrar, Daniel. "Cross-validation." (2019): 542-545.

[15] Greenacre, Michael, et al. "Principal component analysis." Nature Reviews Methods Primers 2.1 (2022): 100.

[16] Sinaga, Kristina P., and Miin-Shen Yang. "Unsupervised K-means clustering algorithm." IEEE access 8 (2020): 80716-80727.

[17] Bushra, Adil Abdu, and Gangman Yi. "Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms." IEEE Access 9 (2021): 87918-87935.

[18] De Diego, I. M., Redondo, A. R., Fernández, R. R., Navarro, J., & Moguerza, J. M. (2022). General performance score for classification problems. Applied Intelligence, 52(10), 12049-12063.

[19] Hancock III, John T., Taghi M. Khoshgoftaar, and Justin M. Johnson. "Using area under the precision recall curve to assess the effect of random undersampling in the classification of imbalanced medicare big data." International Journal of Reliability, Quality and Safety Engineering 31.01 (2024): 2350039.

[20] Salehi, Amir Reza, and Majid Khedmati. "A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data." Scientific Reports 14.1 (2024): 5152.

A Multi-Agent System for Student Banking and Stock Market Analysis

Darko Trajanov
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, North Macedonia
dare.trajanov@gmail.com

***Abstract*—This paper presents a multi-agent AI system for student-focused personal banking and introductory stock market analysis. Two core agents collaborate: one to personalize weekly budgeting goals, savings challenges, and cost-cutting strategies based on user spending, and another to synthesize fundamental and sentiment stock indicators to encourage prudent early investing. Unlike traditional banking chatbots, our system orchestrates specialized sub-agents (e.g., sentiment analysis, data cleansing) via an agent coordination framework and employs a shared-memory structure to synchronize short-term budgeting constraints with trading suggestions. We also incorporate an explainability module to foster user trust and financial literacy. A three-week study with ten university students (and a matched control group using a simplified budgeting chatbot) showed the multi-agent system significantly improved objective saving rates, subjective financial confidence, and engagement with responsible investing behaviors. While still a pilot, this work demonstrates how LLMs, multi-agent orchestration, and XAI can reimagine student banking by bridging day-to-day money management and early-stage investing.**

***Index Terms*—Student Banking, Budgeting, AI Agents, Multi-Agent Systems, Stock Market Analysis, Explainable AI, Shared Memory, CrewAI**

I. Introduction

Students often face rising tuition, housing costs, and consumer debt, often lacking exposure to investing [1]–[3]. Although banking chatbots exist for routine queries, they rarely offer personalized strategies for budgeting or beginner-friendly investment guidance. Consequently, bridging short-term budget management with foundational stock market literacy is a notable gap [4], [5]. This gap not only hinders immediate financial well-being but also misses opportunities for early financial literacy development, potentially impacting long-term economic security.

Recent research shows large language models (LLMs) and multi-agent systems enable adaptive, personalized, explainable financial tools [6]–[10]. Such approaches are vital for younger users needing trust and financial awareness [11], [12]. Key challenges include:

1) Ensuring **explainable** and transparent recommendations, crucial for trust [13].
2) Maintaining a **shared memory** so budgeting informs investment suggestions [14], [15].
3) Incorporating **robust evaluation** using surveys and objective metrics (e.g., savings) [16], [17].

In response, this paper introduces a multi-agent AI system tailored for students, addressing budgeting and basic stock investing recommendations. The system emphasizes:

- An **explainability layer** clarifying how sentiment analysis and fundamental metrics drive decisions [18].
- A **shared-memory** approach linking spending patterns directly to recommended stock transactions [19].
- A **three-week user study** assessing objective savings and self-reported financial confidence vs. a control group.

A. Contributions

Addressing the identified challenges and leveraging our proposed architecture, this paper makes the following specific contributions:

- **Multi-Agent Orchestration for Student Finance**: We detail how CrewAI coordinates sub-agents (sentiment classification, fundamental analysis, budget categorization) unified via an in-memory store [22].
- **Tight Coupling of Budgeting and Investing**: Budget overruns automatically scale down or alter recommended trading actions, encouraging prudent behavior [23].
- **Explainability and User Study**: We embed short rationales (e.g., "Tesla has low P/E, but weekly expense exceeded thresholds") and evaluate effects on savings outcomes and user confidence relative to a control group [21].

The source code for this project can be found on GitHub[1].

II. Related Work

A. Multi-Agent Systems and Financial Applications

Multi-agent systems gain traction in automated portfolio management [8], risk assessment [14], and fraud detection [3]. Frameworks like *AgentVerse* explore emergent behaviors [6], while specialized approaches use multi-agent reinforcement learning for dynamic portfolio optimization [15], [16]. Our approach integrates multi-agent coordination into a student-facing environment, linking budgeting data with basic investing strategies. Using an in-memory shared data store ensures

[1]https://github.com/trajnovd/sb-aiagents

real-time consistency among sub-agents [4], enabling user-centered features like micro-challenges. Unlike purely RL-based market optimizers [15], our system emphasizes the user's daily habits influencing investment capacity.

B. LLMs, Explainability, and Student Finance

LLMs, including financial variants [23], enable personalized financial interactions but risk "black-box" recommendations eroding trust [9], [18]. Explainable AI (XAI) mitigates this by clarifying how metrics (e.g., P/E ratio, sentiment) lead to advice [13], [17].

Facilitating trust for financially inexperienced users like students is key [5]. Effective design involves comprehensible explanations and adaptive architecture [7]. Our study contributes by evaluating how an XAI-driven approach impacts user confidence, savings, and engagement.

III. System Architecture

A. High-Level Overview

Figure 1 illustrates our system. Two primary agents anchor the design:

- **Budgeting & Savings Agent**: Collects transaction data, enforces limits, suggests savings challenges [21].
- **Stock Analysis & Trading Agent**: Integrates sentiment/fundamentals for buy/hold/sell actions.

They operate atop CrewAI's engine, spawning sub-agents (e.g., "sentiment") via API [19], [24]. A shared store records budget states, goals, and portfolio decisions for synchronization. CrewAI manages the workflow by defining specific tasks for each agent and sub-agent, often triggered by user input (e.g., logging expenses) or scheduled events (e.g., weekly review). The shared memory, implemented using an in-memory key-value store analogous to Redis, acts as a central blackboard. Agents publish state updates (e.g., current budget status, sentiment scores) and subscribe to relevant changes, ensuring, for instance, that the Trading Agent always accesses the latest budget constraints before generating recommendations. Simple locking mechanisms prevent race conditions during simultaneous read/write operations.

B. Budgeting and Savings Agent

1) Transaction Tagging: Users label their daily transactions (e.g., groceries, rent, transport), while the agent can auto-suggest categories using merchant metadata and contextual cues [19]. Instances of overspending trigger notifications that are stored in the shared memory, along with suggestions for cost-cutting or budget adjustments.

2) Micro-Challenges and Reports: Each week, the agent proposes personalized savings challenges (e.g., "Cut $10 from transport") based on recent spending behavior. These are generated using prompt-based planning with a GPT-4o model [19], [20]. For example, a prompt might ask the model to suggest a specific, achievable $10–$20 reduction in a non-essential category, framed in a supportive tone. Challenge outcomes update the user's risk profile; for instance, overspending can limit investment recommendations.

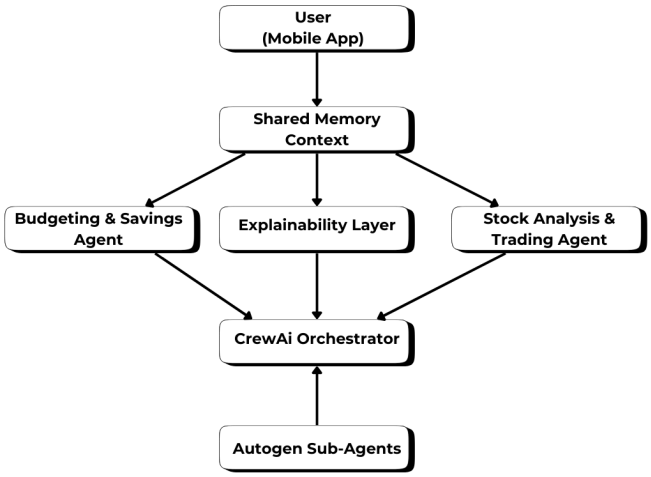


Fig. 1. High-level system diagram featuring CrewAI, sub-agents, and a shared memory store.

Beyond daily feedback, the agent compiles **weekly and monthly reports** summarizing budget performance, category-level spending trends, and progress toward goals. It can also provide **alternative spending suggestions** aligned with user-defined priorities (e.g., "Cook at home instead of eating out") to promote more sustainable financial habits.

C. Stock Analysis and Trading Agent

1) Sentiment Analysis: A dedicated sub-agent retrieves relevant financial news and uses a GPT-4o classifier to compute sentiment scores ranging from -1 (bearish) to +1 (bullish) [21]. Articles with positive earnings reports or strategic developments raise the system's recommendation confidence [7].

2) Fundamental Checks: The agent accesses key financial indicators—Price-to-Earnings (P/E) ratio, Earnings Per Share (EPS) growth, and Debt-to-Equity ratio—via a third-party financial API [8]. These metrics are evaluated using rule-based heuristics to determine whether a stock appears overvalued, fairly priced, or undervalued. Future versions may incorporate knowledge graph-based reasoning for deeper assessments [14].

3) Trading Logic with Shared Memory and Embedded Explanations: The Trading Agent dynamically adjusts its recommendations based on financial behavior tracked by the Budgeting Agent. For example, users flagged for overspending or missed savings goals receive throttled trading suggestions (e.g., smaller position sizes), while users with consistent savings may be offered slightly higher exposure to attractive opportunities [12].

IV. User Study and Results

A. Methodology

We ran a three-week test involving 20 total participants, divided into two groups:

- **Experimental Group (EG),** 10 students using the full multi-agent system (budgeting + investing + integration).

- **Control Group (CG),** 10 students using a simpler budgeting-only chatbot (no investing, no shared memory).

We recruited 20 undergraduate participants aged 19-23 through personal contacts from computer science and economics faculties. All participants confirmed they managed their own finances and had limited prior formal investment experience. All participants were instructed to log all expenditures daily via the provided interface, categorizing them into predefined groups (e.g., food, transport, leisure, bills). Both groups completed pre/post surveys measuring confidence in financial decision-making using a single-item 7-point Likert scale question ('How confident do you feel in making personal financial decisions?', where 1 = 'Not at all confident' and 7 = 'Extremely confident'). The control group interacted with a standard rule-based chatbot focused solely on budget tracking; it allowed expense logging by category and viewing weekly summaries but offered no personalized savings tips, micro-challenges, or investment features, providing a baseline for standard digital budgeting assistance. Experimental group participants additionally received weekly stock recommendations (1-2 stocks, buy/hold/sell) with explanations via the app.They used an integrated paper trading module, simulating a standard virtual brokerage environment, with a starting virtual balance of $100,000 to execute suggested trades if they chose [11].

B. Quantitative Outcomes

TABLE I
SELECTED RESULTS AFTER 3-WEEK TRIAL

Metric	Control Group	Experimental Group
Avg. Weekly Savings Increase	+$14.2	+$44.7
Confidence Gain (Scale 1–7)	+0.5	+1.6
Paper-Trade ROI (%)	–	+3.1

Table I shows EG saved more (+$44.7 vs. +$14.2) and gained more confidence (+1.6 vs +0.5). EG paper-trading averaged +3.1% ROI. Two-sample t-test on savings increase is significant ($p < 0.05$).

C. Time-on-Task and App Usage Frequency

We tracked two engagement metrics: average **time-on-task** and **daily app sessions**. The Experimental Group (EG) averaged **15 minutes/day**, compared to **11.6 minutes/day** for the Control Group (CG). This **notable difference** suggests higher engagement, and interviews indicated the investing component enhanced user interest. Backend logs showed that 60% of EG users opened the app at least twice daily, versus 30% in CG. This indicates that coupling budgeting with simulated investing fosters higher engagement. Future studies may use richer behavioral tracking (e.g., clickstream or heatmaps) to better assess long-term retention.

D. Sentiment Chart Example

Fig. 2 shows sample sentiment scores for five news headlines related to a stock. Positive bars = bullish, negative = bearish/neutral [9].

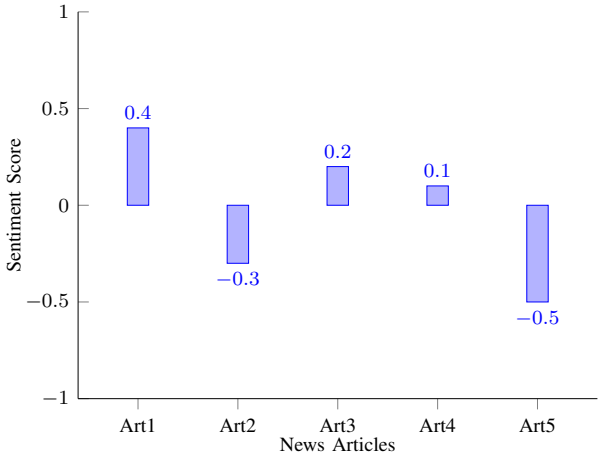


Fig. 2. Sample sentiment scores for five news headlines (Tesla). Positive = bullish, negative = bearish/neutral.

E. Qualitative Feedback

Qualitative interviews revealed nuances: EG participants (8/10) felt daily tagging increased awareness of "unnecessary spending." Many (7/10) appreciated the link between saving efforts and potential investment opportunities, finding it motivating. Several noted the explanations, while simple, helped demystify terms like 'P/E ratio'. Conversely, CG users often described their tool as "functional but dry," lacking features to encourage proactive financial planning or learning, with several explicitly wishing for integrated, personalized advice beyond basic budget summaries.

V. Extended Discussion & Future Work

A. Addressing Reference Breadth

Initial reliance on fewer sources expanded to include peer-reviewed work in multi-agent collaboration [6], LLM planning [19], [20], financial XAI [13], [18], and automated portfolio management [8], [16]. This strengthens the conceptual foundation.

B. Empirical Rigor and Limitations

Our 3-week, 20-participant study improves on earlier pilots, adding basic statistics and engagement metrics. Limitations remain: modest sample size, short duration, paper-trading proxy. Future studies need longer durations, larger samples, potentially real funds/simulators [21]. Further limitations include the reliance on self-reported expense logging, prone to inaccuracies, and focus on basic stocks/indicators, neglecting broader market factors."

C. Framing the Novelty as Systems Application

This work is a *socio-technical integration*, not primarily an algorithmic breakthrough. Contributions lie in:

- **Seamless Orchestration**: Coordinating sub-agents via shared memory/concurrency [6].

- **Socio-Technical Coupling**: Linking real-time budgeting to trading suggestions for responsible behavior [12].
- **Explainable Nudges**: User-friendly rationales building confidence [9].

Framing as applied systems highlights impact bridging student budgeting/investing. This socio-technical perspective is crucial. The system is intended to engage with users and subtly influence their behavior rather than just automating financial duties. Building cognitive trust is the goal of the explainability layer, and acting as a behavioral restraint is the linkage of investment recommendations with budgeting reality. Evaluating success thus requires measuring not just ROI or savings but also user understanding, confidence, and adherence to responsible financial practices over time.

D. Deeper Technical Novelty

Vs. traditional chatbots, our multi-agent orchestration paradigm uses:

- *CrewAI Orchestration/Sub-Agents*: Coordinates specialized agents (sentiment, fundamental, cleansing) for concurrent tasks with shared context [22].
- *Prompting/Fine-Tuning LLMs*: Uses GPT-4o model for domain-specific sentiment [19], [21].
- *Shared Memory*: Redis-like store ensures budget flags update trading logic constraints immediately. Concurrency control prevents inconsistent advice [6].
- *Rule-Based vs. ML*: Combines rule heuristics (P/E thresholds) with risk profiles. ML could refine thresholds [15].

E. Next Steps

Based on our findings and limitations, key directions for future work include:

- **Longer Evaluations/Real Funds**: Extend trials (semester+); consider real-money integration (subject to ethics).
- **Automating Tagging**: Leverage metadata/feedback loops to reduce manual labeling [19].
- **Enhanced Portfolio Logic**: Expand to portfolio metrics (diversification, Sharpe ratios) [8].
- **Adaptive Explanations**: Tailor explanation depth to user literacy [9], [10].

VI. Conclusion

We presented a multi-agent system integrating budgeting feedback and stock analysis via shared memory for student financial guidance. A controlled study showed coupling budgeting/investing advice significantly boosts savings and confidence vs. a basic chatbot. Explainability promotes financial literacy. Future work involves expanding coordination, data pipelines, and validation duration/scale. Advances in agent orchestration [6] promise refined integration of spending, psychology, and market insights. Findings underscore the impact of socio-technical integration: bridging MAS, LLM explanations, and user feedback transforms personal finance management for students.

References

[1] A. Mittal and R. Goel, "Sentiment Analysis in Stock Trading: A Review of Techniques and Applications," *Appl. Artif. Intell.*, vol. 35, no. 14, pp. 1058–1073, 2021.

[2] L. Chen and T. Nguyen, "Impact of AI Chatbots on Online Banking Engagement," *J. Digit. Finance*, vol. 11, no. 2, pp. 145–157, 2023.

[3] J. Klynge, "AI in Banking and Finance: Bridging Consumer Expectations and Institutional Capabilities," *Banking Innov. J.*, vol. 2, no. 1, pp. 22–35, 2023.

[4] M. Lanham, *AI Agents in Action*, Manning Publications, 2025.

[5] A. Fernandez and M. Rodriguez, "Emerging Trends in AI for Financial Literacy and Investment Education," *J. Financ. Mark.*, vol. 56, pp. 101–115, 2023.

[6] C. Yang *et al.*, "AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors in Agents," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, pp. 1–28, 2024.

[7] J. Park *et al.*, "Generative Agents: Interactive Simulacra of Human Behavior," in *Proc. ACM Conf. Hum. Factors Comput. Syst. (CHI)*, 2023, pp. 1–22.

[8] Y. Fu, H. Peng, and Y. Cao, "Financially Intelligent Agents: A Survey of Automated Portfolio Management and Beyond," *J. Finance Data Sci.*, vol. 9, no. 3, p. 100079, 2023.

[9] X. Li, Y. Chen, and J. Zhang, "Explainable Artificial Intelligence for Financial Applications: A Survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 5, pp. 2001–2015, 2022.

[10] A. Kumar, R. Singh, and D. Patel, "A Review of Explainable AI Techniques in Finance," *Expert Syst. Appl.*, vol. 175, p. 114776, 2021.

[11] D. Li and Q. Sun, "Explainable Deep Learning Models for Stock Market Analysis," *Neurocomputing*, vol. 450, pp. 254–265, 2021.

[12] S. Patel and L. Zhao, "Hybrid Multi-Agent Systems for Enhancing Financial Decision-Making: A Case Study on Student Banking," *IEEE Access*, vol. 10, pp. 9023–9035, 2022.

[13] L. Salewski, A. Moos, and S. Friederich, "Causal and Explainable AI for Financial Technology," *Digit. Finance*, vol. 5, no. 1, pp. 33–55, 2023.

[14] J. Zhan *et al.*, "Enhancing Knowledge Graph Reasoning for Financial Risk Assessment: A Multi-Agent Approach," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 1, pp. 453–466, 2024.

[15] Y. Wang, K. Xu, and Y. Huang, "Multi-Agent Reinforcement Learning in Financial Markets: Methods and Applications," *IEEE Access*, vol. 10, pp. 8000–8015, 2022.

[16] H. Park and S. Kim, "Dynamic Portfolio Optimization Using Multi-Agent Deep Reinforcement Learning," *IEEE Trans. Comput. Intell. AI Games*, vol. 14, no. 2, pp. 205–217, 2022.

[17] J. Ouyang and W. Li, "Integrating Multi-Agent Systems with Explainable AI for Automated Financial Advisory," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 3, pp. 1280–1292, 2022.

[18] R. Gupta and P. Mehta, "Explainable Multi-Agent Systems in Fintech: Bridging the Gap Between AI and User Trust," *Inf. Fusion*, vol. 70, pp. 45–59, 2023.

[19] S. Li *et al.*, "Pre-trained Language Models as Visual Planners for Human Assistance," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 54612–54628.

[20] K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati, "Large Language Models as Commonsense Knowledge for Large-Scale Task Planning," *IEEE Robot. Autom. Lett.*, vol. 8, no. 11, pp. 7032–7039, 2023.

[21] L. Chen and X. Zhao, "Recent Advances in AI-Driven Financial Decision Support Systems," *J. Financ. Data Sci.*, vol. 5, no. 1, pp. 45–62, 2023.

[22] J. Zheng, J. B. Tenenbaum, and I. Sucholutsky, "Agent-based modeling with large language models," *Nat. Mach. Intell.*, vol. 5, pp. 1052–1063, 2023.

[23] P. Vaswani, G. Kasneci, and W. Wang, "FinGPT: Open-Source Financial Large Language Models," *arXiv preprint arXiv:2306.06031*, 2023. Presented at *4th ACM Int. Conf. AI Finance (ICAIF)*, 2023.

[24] CrewAI Documentation, "Developer's Guide." [Online]. Available: https://docs.crew.ai/. [Accessed: 2025-04-01].

Going NUTS with ADVI: Exploring various Bayesian inference techniques with Facebook Prophet

Jovan Krajevski
Faculty of Computer Science and Engineering
University "Ss. Cyril and Methodius"
Skopje, R.N. Macedonia
jovan.krajevski@gmail.com

Biljana Tojtovska Ribarski
Faculty of Computer Science and Engineering
University "Ss. Cyril and Methodius"
Skopje, R.N. Macedonia
biljana.tojtovska@finki.ukim.mk

***Abstract*—Since its introduction, Facebook Prophet has attracted positive attention from both classical statisticians and the Bayesian statistics community. The model provides two built-in inference methods: maximum a posteriori estimation using the L-BFGS-B algorithm, and Markov Chain Monte Carlo (MCMC) sampling via the No-U-Turn Sampler (NUTS). While exploring various time-series forecasting problems using Bayesian inference with Prophet, we encountered limitations stemming from the inability to apply alternative inference techniques beyond those provided by default. Additionally, the fluent API design of Facebook Prophet proved insufficiently flexible for implementing our custom modeling ideas. To address these shortcomings, we developed a complete reimplementation of the Prophet model in PyMC, which enables us to extend the base model and evaluate and compare multiple Bayesian inference methods. In this paper, we present our PyMC-based implementation and analyze in detail the implementation of different Bayesian inference techniques. We consider full MCMC techniques, MAP estimation and Variational inference techniques on a time-series forecasting problem. We discuss in details the sampling approach, convergence diagnostics, forecasting metrics as well as their computational efficiency and detect possible issues which will be addressed in our future work.**

***Index Terms*—Bayesian inference, Facebook Prophet, Variational Inference, Markov chain Monte Carlo, maximum a posteriori estimate.**

I. Introduction

Performing full Bayesian inference has always posed significant computational challenges. While the statistical quality obtained by sampling from the true posterior using Markov Chain Monte Carlo (MCMC) methods is widely recognized and rarely questioned, the computational cost involved often raises doubts about whether it is worthwhile to perform sampling. On the other hand, maximum a posteriori (MAP) estimates are computationally efficient but they collapse the full posterior distribution into a single point estimate, discarding information about parameter uncertainty that is central to Bayesian analysis.

Variational Inference (VI) is a middle-ground approach that bridges the gap between computationally demanding MCMC sampling and overly simplistic MAP estimation. By providing an analytical approximation to the true posterior and sampling from it, VI significantly reduces computational overhead compared to full MCMC sampling. However, like many other approximation techniques, it does not always capture all essential characteristics of the posterior.

There is no universally optimal Bayesian inference method - the choice of an optimal method depends on the model complexity, data size and inferential goals. The goal of this paper is to study the forecasting performance and computational accuracy of different Bayesian inference methods. We analyze the application of full MCMC sampling, VI, and MAP estimation and their various algorithmic implementations on a specific model and forecasting scenario. This investigation arose during our exploration of Bayesian transfer learning techniques for time-series forecasting tasks using Facebook Prophet [1]. Throughout our research, we quickly encountered limitations related to Prophet's fluent API design [2], which restricted our ability to incorporate new components into Prophet's decomposable time-series model [3]. Additionally, Prophet's built-in inference methods proved insufficient: full MCMC sampling was computationally prohibitive for rapid experimentation, while MAP estimation alone did not adequately represent parameter uncertainty. To overcome these limitations, we developed a complete reimplementation of Facebook Prophet using PyMC [4].

In this paper, we first go over our PyMC-based reimplementation. We give a quick overview of the FP model, before introducing the new intuitive API that we will implement. Section II also includes theoretical introduction of various Bayesian inference techniques. The methodology for their comparison and the results are presented in Section III and IV. The conclusions and future work are given in Section V and VI.

II. Reimplementing Facebook Prophet in PyMC

A. Overview of the model

Facebook Prophet is a decomposable time series model given by

This paper was partially supported by Faculty of Computer Science and Engineering, Skopje, R.N. Macedonia

$$y(t) \sim g(t) + \sum_{i=1}^{M} s(t; p_i, c_i) + h(t) + \epsilon(t) \quad (1)$$

where $y(t)$ is the time-series that is being modeled, $g(t)$ is a trend component, $s(t; p_i, c_i)$ is a Fourier seasonality component with a period of p_i represented with a series of order c_i, $h(t)$ is a holiday component and $\epsilon(t)$ is normally distributed random error [1].

Facebook Prophet allows the user to specify additional regressors. However, in this paper and in our future research, we work with univariate time-series, so additional regressors will be ignored.

Facebook Prophet gives an option of three different choices for the trend component $g(t)$: a piece-wise linear trend, a piece-wise logistic trend and a flat trend. In the piece-wise trends, it is up to the user to define the number of change points that will be taken into consideration. Each new change point introduces a new parameter that needs to be estimated, increasing the computational complexity of the model. By default, Facebook Prophet uses 25 change points.

The user is allowed to specify the series order of the Fourier seasonality component $s(t; p_i, c_i)$, i.e. the number of sine and cosine terms that will be used to approximate the periodical seasonality. Increasing the series order by one introduces one additional sine and one additional cosine term, by which the number of parameters that need to be estimated is increased by two. By default, Facebook Prophet fits two seasonality components: a yearly seasonality (period of 365.25) approximated with a Fourier series of order 10, and a weekly seasonality (period of 7) approximated with a Fourier series of order 3.

In the rest of this paper, the holiday component will be removed since it is not significant for our problem.

B. Intuitive API

Facebook Prophet's current implementation lacks the flexibility needed to introduce new components and only allows the user to do NUTS sampling or perform a MAP estimate. Its API employs a fluent interface, which is also less intuitive than the mathematical notation used to specify the model.

In our reimplementation of Facebook Prophet, we guide the development of the API by following other reimplementations in PyMC [5]–[7]. The resulting API, inspired by TimeSeers [8], mirrors the mathematical notation typically used when writing models.

The main idea behind this API is to be modular - the user is allowed to define each component separately. So, if the user wishes to define a linear trend component with 10 changepoints, they would do that by instantiating an object from a `LinearTrend` component by simply writing `trend = LinearTrend(n_changepoints=10)`.

After defining the components, constructing the model is straightforward, because the API implements the summation and multiplication operators for these objects. If the user defined a linear trend `trend`, a yearly seasonality `yearly` and a weekly seasonality `weekly`, they can create a multiplicative seasonality model by simply writing `model = trend * (1 + yearly + weekly)`.

Below, we present the same Prophet model expressed in the fluent API from Facebook Prophet and in the more intuitive API we used in our reimplementation.

Listing 1. Two different APIs for specifying Prophet-like models

```
#fluent API from Facebook Prophet
model = Prophet(
    seasonality_mode="multiplicative",
    yearly_seasonality=10,
    weekly_seasonality=3,
).add_seasonality(
    name="monthly",
    period=30.5,
    fourier_order=5,
)

#new intuitive API
model = LinearTrend() * (
    1
    + FourierSeasonality(p=365.25, order=10)
    + FourierSeasonality(p=30.5, order=5)
    + FourierSeasonality(p=7, order=3)
)
```

After introducing the new API, we briefly go over the details of full MCMC sampling, MAP estimation and VI.

C. Full MCMC sampling

We investigate three different algorithms that perform MCMC sampling: Metropolis-Hastings (MH) [9], [10], NUTS [11] and Z-score adaptation of DEMetropolis (DMZ) [12]. These algorithms differ essentially. MH is the original random-walk approach in sampling from the posterior. The NUTS algorithm is an extension of the Hamiltonian Monte Carlo (HMC) - it requires computation of the gradient of the log posterior, which points to the high-density regions of the posterior. On the other hand DMZ is a population-based, gradient-free MCMC algorithm that uses differential evolution between chains in a standardized Z-space as its proposal mechanism.

All of these approaches generate samples that approximate the true posterior distribution. However, due to limitations in the sampling algorithms and the inherent stochasticity of MCMC, we typically require a large number of samples to obtain reliable statistical estimates. MCMC sampling usually includes a warm-up (or adaptation) phase, during which the algorithm adjusts its internal parameters and moves away from the initial starting point toward areas of higher posterior density. This process is typically performed across multiple independent chains. Warm-up samples are generally discarded before inference begins.

D. MAP estimate

The MAP estimate equals the mode of the posterior distribution. It is Bayesian regularized generalization of maximum likelihood estimation (MLE) - the maximum likelihood

[1]This case is known as "additive seasonality". If the seasonality is multiplicative, (1) becomes $y(t) \sim g(t)\left(1 + \sum_{i=1}^{M} s(t; p_i, c_i) + h(t)\right) + \epsilon(t)$.

function is modified by including a prior distribution over the parameters, effectively regularizing the solution according to prior knowledge. Since this is a point estimate, it fails to capture all of the characteristics of the posterior and closely resembles the philosophy behind frequentist inference. For example, given a multi-modal posterior, the MAP estimate will not inform us of the existence of the other ”modes”.

In this paper, we will investigate performin MAP using L-BFGS-B [13]. It is a fast, gradient based (quasi-Newton) optimization algorithm which is memory efficient in presence of large number of parameters and thus suitable for MAP estimation.

E. Variational Inference

In VI, the true posterior is approximated by a variational distribution with a known analytical form, usually by minimizing the Kullback–Leibler divergence between the true posterior and the variational distribution. The sampling is performed from the variational distribution, which is a faster process than sampling from the true posterior.

In this paper, we investigate two different VI algorithms: ADVI [14] and FullRank ADVI (FR ADVI) [15] ADVI employs a factorized Gaussian approximation for computational efficiency, while FR ADVI enhances this by modeling parameter correlations through a full-rank Gaussian. For more details we refer to the existing literature.

III. Methodology

A. Dataset

For analysis of the introduced techniques we choose a time series of the log daily page views for the Wikipedia page of Peyton Manning, available at the Facebook Prophet documentation and GitHub repository. We use two years of data to train the model and we forecast the next year.

B. Model specification

We use the default model offered by Facebook Prophet, without doing any hyperparameter tuning. This model uses a piecewise linear trend with 25 change points and additive seasonality that takes into account yearly and weekly seasonal patterns.

In Table I, we present the parameters that we are sampling from. For each parameter, we denote the shape i.e. the number of independent random variables with the same prior.

TABLE I
Parameters that are being sampled in the Facebook Prophet model

Component	Parameter	Shape	Prior
Linear Trend - $g(t)$	slope - k	1	$Normal(0, 5)$
	intercept - m	1	$Normal(0, 5)$
	change points - δ	25	$Laplace(0, 0.05)$
Yearly Seasonality - $s(t; 365.25, 10)$	Fourier coefficients - β_y	$2 \cdot 10$	$Normal(0, 10)$
Weekly Seasonality - $s(t; 7, 3)$	Fourier coefficients - β_w	$2 \cdot 3$	$Normal(0, 10)$

C. Sampling, Convergence Diagnostics and Forecasting Metrics

When using the NUTS method, we drew 4 chains of 2000 samples each with a warm-up of 1000 samples that we discarded. We ran MH and DMZ with 1000000 samples, because the Effective Sample Size was too low, whereas the R-hat Diagnostic was too high for a smaller amount of samples. The warm-up for MH and DMZ consisted of 10000 samples that were discarded. For the VI techniques, we produced 2000 samples after running 100000 iterations.

We additionally included convergence diagnostics. For the full MCMC sampling, we analyzed and present the plot of the posterior samples, the autocorrelation plot of the posterior samples, the Effective Sample Size (ESS) and the R-hat (Gelman-Rubin) diagnostics. For the ADVI and FR ADVI techniques, we analyzed and present the negative Evidence Lower Bound (ELBO) convergence.

For each of the different inference methods, we calculated and compared the standard forecasting metrics (MAE, MSE, RMSE, MAPE) and the execution time. For those methods that reached convergence, we also present their forecasting metrics and execution time at the moment they reached convergence.

The simulations were executed on a machine with AMD 7950x processor with 64GB DDR5 system RAM running Debian 12.

IV. Results

A. MCMC convergence

1) R-hat: The R-hat diagnostic statistic compares the between-chain and in-chain variances. R-hat value larger than 1.01 is evidence of a lack of convergence of the sampling algorithm. In Table II, where the R-hat diagnostic is calculated for each parameter from Table I, we can see that only the NUTS sampling algorithm produces chains that converge.

TABLE II
Comparing the R-hat diagnostic for each parameter

Method	k	m	δ	β_y	β_w
MH	1.0443	1.0249	1.0032	1.0010	1.0000
DMZ	1.4047	1.4168	1.5264	1.1215	1.0205
NUTS	1.0005	1.0010	1.0010	1.0004	1.0004

2) Efective Sample Size: The ESS for an MCMC method quantifies the number of “independent” samples in the chain, taking into account the autocorrelation between successive samples. Higher ESS implies that the estimates based on the chain (as means, variances etc.) are more reliable. We adopt $ESS > 400$ as the threshold for sufficient independent samples to ensure stable posterior estimates.

Drawing 1000000 samples with MH and DMZ did not yield an ESS larger than 400 for each parameter. The values are presented in Table III. The small number of independent samples suggests that the chains are not exploring the posterior distribution efficiently. ESS values for NUTS for all parameters are significantly higher and this was achieved after drawing 2000 samples.

TABLE III
COMPARING THE ESS FOR EACH PARAMETER

Method	k	m	δ	β_y	β_w
MH	82.78	163.80	5697.00	274152.10	795493.71
DMZ	8.66	8.52	9.70	86.80	328.91
NUTS	3050.72	3267.21	7033.49	9089.82	10440.54

3) Autocorrelation in MCMC methods: The family of MCMC algorithms that we consider makes an essential assumption - every new sample accepted in the chain depends only on the last sample added to the chain. Thus autocorrelation inherently exists in the produced samples.

We show the MH plot of the slope of the linear trend in Fig. 1, noting that the DMZ plot looks similar. The other parameters behave the same, with a remark that the parameters that define the change points and seasonalities have slightly better autocorrelation plots.

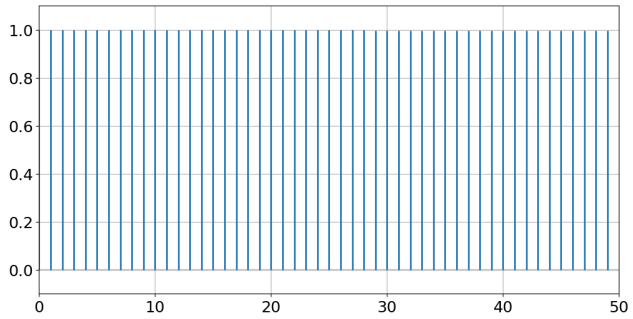


Fig. 1. Autocorrelation plot from Metropolis-Hastings for the slope parameter k in the Linear Trend component

The autocorrelation plots of the posterior samples confirm what we found out from the ESS comparison. MH and DMZ generate autocorrelated chains that do not explore the posterior efficiently. Even with a lag of 50, the sample from MH shows a strong autocorrelation. NUTS, on the other hand, starts off with a much smaller autocorrelation for lags smaller than 5, and the autocorrelation gradually vanishes, as can be seen in Fig. 2.

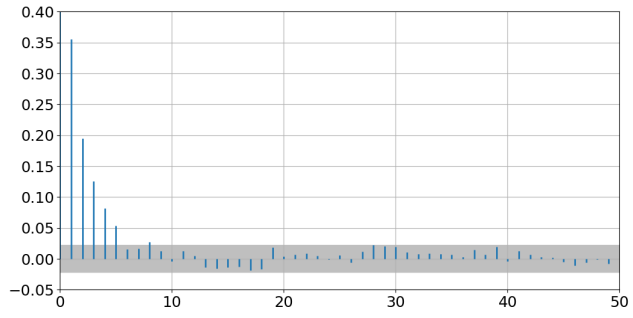


Fig. 2. Autocorrelation plot from NUTS for the slope parameter k in the Linear Trend component

The presented diagnostics show that NUTS converges well for all parameters. To improve the MH and DMZ methods, we may need to draw more samples, run more chains, tune the sampler or adjust the proposal distribution or step sizes. But even if we generate a long chain of samples, still the amount of independent information may be low, due to the correlation of the samples.

B. VI convergence

In the VI methods, the Evidence Lower Bound (ELBO) is the objective function that is maximized in order to find the best approximation to the true posterior. When the loss function (negative ELBO) has reached a steady state, it indicates that the variational inference procedure has effectively converged. However, this does not guarantee a high-quality posterior approximation. We need to monitor the ELBO curve and perform other diagnostic checks. To determine when the negative ELBO reaches a plateau, we plot the negative ELBO for both ADVI and FR ADVI through the iterations in Fig. 3.

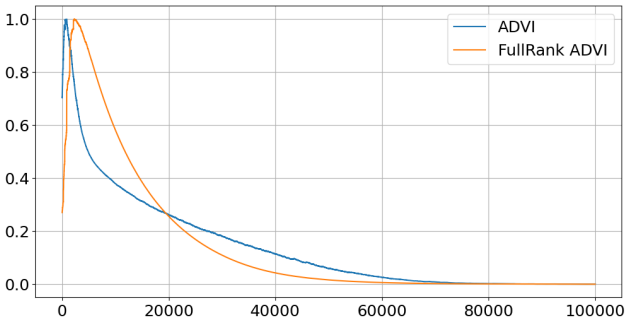


Fig. 3. Negative ELBO through iterations

To reduce noise, the plots are smoothed using exponential smoothing with a coefficent of 10^{-3} for ADVI and 10^{-4} for FR ADVI. Additionally, they are scaled in the $[0, 1]$ range because their ELBOs differ in absolute value, making it more difficult to notice the convergence. We observe that convergence is achieved around the 80000-th iteration for both approaches.

C. Forecasting metrics

Table IV presents the standard forecasting metrics for the different MCMC inference techniques after drawing all samples and at their point of convergence.

TABLE IV
COMPARING MCMC FORECASTING METRICS AFTER DRAWING ALL SAMPLES AND AT POINT OF CONVERGENCE

Method	MSE	RMSE	MAE	MAPE
MH	0.6886	0.8298	0.6903	0.0862
DMZ	0.5789	0.7609	0.6180	0.0770
NUTS	0.6720	0.8198	0.6792	0.0848
NUTS at convergence	0.6584	0.8114	0.6698	0.0836

Since MH and DMZ never converge, we only show the metrics at the point of convergence for NUTS. We observe that

there is no significant difference at the point of convergence and after drawing 2000 samples for NUTS.

Interestingly enough, eventough the DMZ method never converges, it still generates forecasts with the best quality.

Table V presents the standard forecasting metrics for the different VI inference techniques after 100000 iterations and at their point of convergence. Increasing the number of iterations in the VI methods improves the metrics after reaching the point of convergence. ADVI generates forecasts with the highest quality.

TABLE V
COMPARING VI FORECASTING METRICS

Method	MSE	RMSE	MAE	MAPE
FR ADVI	0.7966	0.8925	0.7572	0.0947
ADVI	0.7092	0.8421	0.7062	0.0882
FR ADVI at conv.	1.1355	1.0656	0.9325	0.1168
ADVI at convergence	1.2841	1.1332	1.0093	0.1265

Finally, in Table VI, we present the metrics obtained with the MAP method.

TABLE VI
METRICS FROM THE MAP METHOD

Method	MSE	RMSE	MAE	MAPE
MAP	0.6915	0.8316	0.6925	0.0865

D. Plot of the posterior samples

Next, we see what the histograms of the samples drawn from the posteriors with different techniques look like. We are looking at the posterior samples drawn from the slope parameter.

Fig. 4 shows that the sample drawn from the FullRank ADVI approximation is less ”certain” about the value of the slope, i.e. it has a higher standard deviation than the sample drawn from the true posterior via the NUTS algorithm. On the other hand, the sample drawn from the ADVI approximation is overconfident about the value of the slope, i.e. it has a lower standard deviation than the sample drawn from the true posterior via the NUTS algorithm.

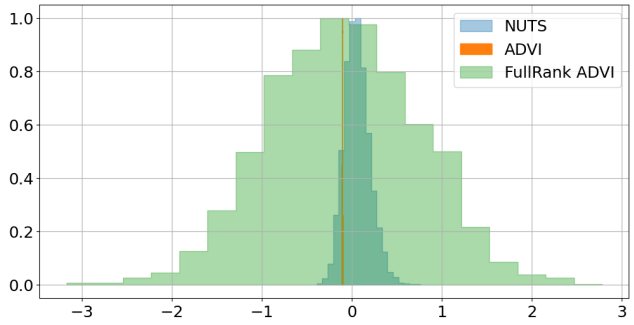


Fig. 4. NUTS vs ADVI and FullRank ADVI for the slope parameter k in the Linear Trend component

Fig. 5 shows that the sample drawn via the MH method is similar with the sample drawn via the NUTS method, but the DMZ method produces a slightly less confident sample.

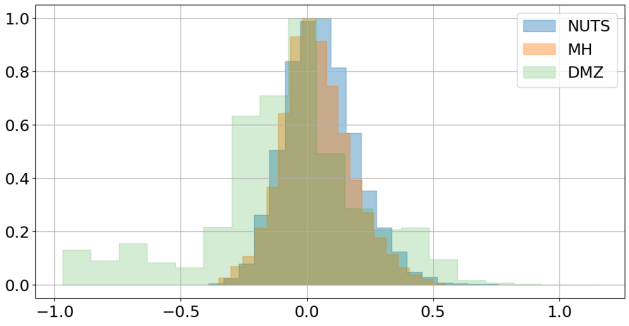


Fig. 5. NUTS vs MH vs DMZ for the slope parameter k in the Linear Trend component

E. Summarized results

Finally, we summarize the performance of the various Bayesian inference methods. Table VII shows the number of samples drawn and the time required for those samples to be drawn for each MCMC. We also present the number of samples and the time required for these methods to converge, noting that only NUTS successfully converges. The last column shows the estimated number of effective samples drawn per second - effective sampling rate is a measure of computational efficiency.

TABLE VII
SUMMARIZED RESULTS FOR THE MCMC METHODS

	MH	DMZ	NUTS
# Samples	1000000	1000000	2000
# Samples to convergance	/	/	550
Time (s)	7141.66	387.13	193.77
Time (s) to convergance	/	/	87.04
ESS per s	150.61	1.14	169.69

NUTS achieves convergence i.e. achieves an R-hat diagnostic statistic less than 1.01 and achives an ESS larger than 400, after only drawing 550 samples. This means that NUTS is very effective, with at least 17% of the samples drawn being effective samples.

We present a similar summarization in Table VIII for the VI methods, only this time we focus on the time required for the total number of iterations of the models and the time required for the models to converge.

TABLE VIII
SUMMARIZED RESULTS FOR THE VI METHODS

	FR ADVI	ADVI
# Iterations	100000	100000
# Iterations to convergance	≈80000	≈80000
Time (s)	22.97	18.25
Time (s) to convergance	18.85	15.06

V. Conclusion

In this paper we analised three major Bayesian inference classes: MAP estimation, full MCMC posterior sampling, and VI on a time-series forecasting problem using the Facebook Prophet model.

MAP estimation consistently delivered high-quality forecasts and benefits from fast execution times. Therefore, MAP is particularly suitable in scenarios where accurately capturing posterior shape and uncertainty is less critical. However, it is important to note that MAP estimates the mode of the posterior distribution, which may differ from the mean usually used in forecasting via MCMC/VI.

MCMC sampling generally provided higher-quality posterior samples compared to VI approaches. This is not surprising since we had to draw large number of samples for MH and DMZ and these methods explore the true posterior. Still, there are some convergence issues that have to be addressed.

Given the computational cost associated with full MCMC sampling, we may consider using ADVI or FullRank ADVI. The forecasting metrics achieved with VI methods are comparable with those of the full MCMC sampling algorithms, but the execution time is reduced significantly. However the FullRank ADVI overestimated and ADVI underestimated posterior uncertainty. This could be due to algorithmic limitations and should be further addressed. Thus NUTS maybe be the best choice for our problem, given the current settings. Its time to convergence is higher compared to VI methods, but not as high as the other MCMC methods. Additionally NUTS showed overall good results for all conducted analyses.

We give a slightly different interpretation of the results that offers some intuition on when to use each technique. We already stated that the MAP estimation is equal to performing MLE with regularization. We can go even further, and state that in the case of Laplace or Normal priors, MAP estimation is equal to performing MLE with L1 or L2 regulatization, respectively (for a Laplace prior with scale b, the regularization strength is $\lambda = \frac{1}{b}$, and for a Normal prior with scale σ, the regularization strength is $\lambda = \frac{1}{2\sigma^2}$).

Then, if MCMC is not fast enough, VI can be used to approximate the uncertainty in these parameters with Laplace and Normal priors. This uncertainty approximation can later be used to determine the regularization strength in both Frequentist and Bayesian frameworks.

VI. Future work

As far as the research of various Bayesian inference techniques goes, in the future we will be looking into GPU acceleration to speed up the full MCMC sampling. Packages like JAX [16] and NumPyro [17] offer GPU acceleration, and they integrate well with PyMC, showing improvements in execution time on large datasets with large models, when the computation of the likelihood is expensive.

However, we need to note that every MCMC algorithm is inherently sequential because of the underlying Markov assumption. In other words, to produce the i-th sample from the posterior, one must produce the $(i-1)$-th sample first and there is no approach that can produce these samples in parallel. Thus, GPU acceleration can only speed up the computation of the likelihood. Even then, if the likelihood is simple, the model does not involve a lot of parameters or the dataset is small, GPU acceleration may end up adding a computational overhead that will overcome the benefits of fast matrix operations, leading to an overall slower sampling approach than a CPU based one.

VI, on the other hand, does not have this sort of limitation. That means that most of our future research will consist of optimizing the VI approximation process by utilizing GPU acceleration. One additional thing worth investigating is the overestimation and underestimation of the posterior uncertainty with FullRank ADVI and ADVI respectively.

Finally, we want to further investigate the usage of Bayesian inference to determine the regularization strength of various parameters, as was mentioned in section V.

References

[1] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[2] M. Fowler and E. Evans, "Fluent interface," *martinfowler. com*, 2005.

[3] A. C. Harvey and S. Peters, "Estimation procedures for structural time series models," *Journal of forecasting*, vol. 9, no. 2, pp. 89–108, 1990.

[4] O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. J. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin *et al.*, "Pymc: a modern, and comprehensive probabilistic programming framework in python," *PeerJ Computer Science*, vol. 9, p. e1516, 2023.

[5] R. Vink, "Build facebook's prophet in pymc3; bayesian time series analyis with generalized additive models," https://www.ritchievink.com/blog/2018/10/09/build-facebooks-prophet-in-pymc3-bayesian-time-series-analyis-with-generalized-additive-models/, 2018.

[6] L. Giacomel, "Pm-prophet," https://github.com/luke14free/pm-prophet, 2020.

[7] D. P. Marco Gorelli, "Air passengers - prophet-like model," in *PyMC examples*, P. Team, Ed.

[8] M. Brouns, "Timeseers," https://github.com/MBrouns/timeseers, 2021.

[9] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[10] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," 1970.

[11] M. D. Hoffman, A. Gelman *et al.*, "The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo." *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1593–1623, 2014.

[12] C. J. Ter Braak and J. A. Vrugt, "Differential evolution markov chain with snooker updater and fewer chains," *Statistics and Computing*, vol. 18, pp. 435–446, 2008.

[13] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on scientific computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[14] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei, "Automatic differentiation variational inference," *Journal of machine learning research*, vol. 18, no. 14, pp. 1–45, 2017.

[15] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, p. 859–877, Apr. 2017. [Online]. Available: http://dx.doi.org/10.1080/01621459.2017.1285773

[16] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs," 2018. [Online]. Available: http://github.com/jax-ml/jax

[17] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman, "Pyro: Deep universal probabilistic programming," *J. Mach. Learn. Res.*, vol. 20, pp. 28:1–28:6, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-403.html

Neural Radiance Fields (NeRFs) for Satellite Imagery: A Literature Review

Arnen Telkjiu
Faculty of Information Sciences and Computer Engineering,
Ss. Cyril and Methodius University in Skopje
Republic of North Macedonia
arnen.telkjiu@students.finki.ukim.mk

Dejan Gjorgjevikj
Faculty of Information Sciences and Computer Engineering,
Ss. Cyril and Methodius University in Skopje
Republic of North Macedonia
dejan.gjorgjevikj@finki.ukim.mk

***Abstract*— This paper presents a brief literature review on the application of Neural Radiance Fields (NeRFs) to satellite imagery for 3D reconstruction. The study outlines the evolution from traditional photogrammetry-based digital surface model generation to deep learning approaches that represent complex scenes as continuous volumetric functions. In particular, the review focuses on two fundamental models, Shadow-NeRF and Satellite-NeRF, which extend the vanilla NeRF framework by incorporating solar illumination metadata and mechanisms to address challenges such as shadows, transient objects, multi-date variations, and data sparsity. A categorization of recent NeRF variants is provided based on their approaches to overcome these challenges, supported by analyses using benchmark satellite datasets. The paper discusses the advantages and limitations of current methodologies and suggests directions for future research in enhancing the accuracy and efficiency of satellite-based 3D reconstructions.**

Keywords*— *neural radiance fields, NeRF, computer vision, machine vision, image processing, satellite imaging, machine learning, neural networks, digital surface models, 3D modelling

I. Introduction

In the 1920s, writing under the pen name of Hermann Noordung, the Slovenian officer Herman Potočnik expressed his ideas on matters surrounding rocketry and space travel in his famous book The Problem of Space Travel. Of the many things he talked about, he even envisioned the observation and research of Earth's surface 'down to the smallest details' for the most varied of reasons [1]. Now, after 100 years, his vision has more than materialized into reality – for indeed, there are a thousand satellites controlled by over 60 countries, that are actively orbiting the Earth for purposes of Earth observation [2]. Apart from being utilized for the bases case of mapping and monitoring, the imagery obtained from these observation satellites has found ample use case in urban planning [3][4], geology [5], navigation [6], meteorology [7], emergency rescue [8][9], agriculture [10], forestry [11], biodiversity conservation [12] and so on. With the rise of the number of satellites and the diversification of their use cases, has also come the rise in data related to EO satellite imagery. By the very nature of satellite imagery, the images come in different dates and different angles. Additionally, these images are subject to variations because of different reasons, from the presence of transient objects (such as cars) to the nature of light and seasonal changes. As regards the purposes of our paper, this changing nature of satellite imagery also affects the 3D reconstruction of the surface these images capture. Put simply, 3D reconstruction in computer vision and visual computing is the task of constructing a 3D model of scanned objects using a set of 2D images [13].

In 3D reconstruction, the large-scale 3D models are usually represented using discrete point clouds or digital surface models (DSMs) of a certain resolution. Because the quality of the DSMs is crucial in this respect, a vast amount of research has been undertaken in the last few decades to enhance their precision and fidelity [14][15]. It is indeed important to note that the algorithms such as semi-global stereo matching (SGM) are considered state-of-the-art in this field, being one of the ways of generating detailed surface maps of urban and natural environments [16]. Stereo-matching algorithms like SGMs firstly generate depth maps and then they fuse these depth maps to create DSMs. The entire 3D reconstruction pipeline involves several steps such as stereo pair selection, dense stereo matching, triangulation, and depth map fusion [17]. But as we said, because of the changing nature of satellite imagery, when processing multi-view satellite images with appearance variations, this approach may struggle to extract effective visual features, resulting in holes or complete failure of the 3D reconstruction pipeline. However, the latest research in 3D modelling from multiple views shows that it is possible to achieve a superior representation of 3D objects and scenes by learning them as continuous functions or fields F [18]. Unlike the traditional methods, NeRF leverages many views to learn to represent the scene as a continuous volumetric representation (i.e., 3D radiance field). Because this representation is uniquely defined by a neural network and can incorporate different aspects of the physical scene like surface radiance or illumination sources – it is seen as a unique way to solve the dense image correspondence problem, as mentioned above [14].

II. Related Work

A. Classical approaches

As we mentioned above, the task of generating DSMs is crucial aspect of 3D reconstruction of Earth's surface. Digital Surface Models (DSMs) represent the topography of the earth's surface, including all objects on it, such as trees, buildings, and other features. They are usually contrasted with Digital Terrain Model (DTMs), which represent the bare ground surface without objects like trees and buildings [19].

In opposition to DTMs, DSMs are a more realistic representation of the surface – capturing both the environment's natural and artificial/man-made features. DSMs can be generated using data from satellite imagery, aerial photography, LiDAR, SAR (synthetic aperture radars) and even drone

surveys. Many papers have been and continues to be written on DSMs generated by optical imagery, because they are usually cheaper than other sensors and have the ability to provide a realistic and detailed representation of the surface for all its worth [20].

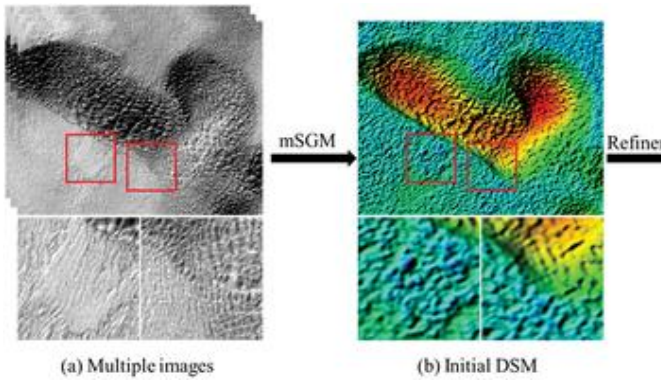


Fig. 1. A simplified pipeline of DSM generation. [21]

Most of the current methods for generating DSMs are based on photogrammetry and involve acquiring the corresponding points in airborne images before relying on image-matching techniques to complete the generation of the DSM.

There are two trends which are at the background of the motivation of our thesis. Remarkable progress has been made in dense geometry 3D reconstruction from aerial images. However, the satellite domain has been much less studied than the traditional pinhole camera model, probably because of the limited availability of high-resolution data. Along with the increased availability of satellite data, the use of satellite data for 3D reconstruction and other aspects is gaining more and more attention [22].

The other trend that is booming in the DSM generation area is the application of deep learning-based algorithms. It is known that currently deep learning-based algorithms are state-of-the-art in this field, but many of these approaches are dependent on supervised learning methods and require ground truth data for training, which is still measured with LiDAR. This data acquisition is expensive, and the quality of the ground truth depends on the density of the generated point cloud. Despite this issue, learning models have the advantage of being trained on a subset of data and tested on many other samples, so the ground truth is just required for the training step, allowing the model to predict in many unseen samples. While the algorithms achieve in general a good reconstruction, their performance can be even improved by finetuning in some samples of the target dataset to reduce the domain gap (if any).

B. NeRF approaches

So the confluence of these two trends – one the one hand, the rise in satellite imagery-based DSM generation, and on the other hand, the increasingly important dependence of DSM generating algorithms on the deep learning world – it is only natural to explore one of the most promising novel technologies in deep learning and 3D reconstruction – which are NeRFs, and especially their implementation in satellite imagery reconstruction pipelines. This new approach, which differs from other learning-based methods, operates in a self-supervised manner. It works on single pixels rather than patches, treats non-Lambertian surfaces and generates new synthetic views [16]. NeRFs have demonstrated impressive capabilities of creating accurate representations of 3D objects or scenes by representing them as a 5D-vector valued function or field. This function implicitly handles tasks such as synthesizing new views and reconstructing 3D information, matching geometry and colors with the camera projections from various angles [23]. NeRFs are good for multi-date as well as multi-view satellite images [17].

In this paper, we have identified and collected a dozen of papers, which amount to a dozen models. In this regard, we have identified two fundamental models, namely FM1: Shadow-NeRF (S-NeRF) and FM2: Satellite-NeRF (Sat-NeRF) and the rest we've described in the table below. We'll discuss the models in more detail in the following sections.

TABLE I. NERF X SATELLITE MODELS

NeRFxSatellite Model	Name
FM1 (SI3)	Shadow-NeRF (S-NeRF)
FM2 (SI4)	Satellite-NeRF (Sat-NeRF)
SI1	Surf-NeRF
SI2	SAT-NGP
SI5	EO-NeRF
SI6	Planet-NeRF
SI7	BRDF-NeRF
SI8	Sparsesat-NeRF
SI9	SatelliteRF
SI10	Season-NeRF
SI11	Sat-Mesh

III. METHODOLOGY

A. From Vanilla NeRF to NeRFxSatellite: The 5 Problems

Satellite imagery, although at a highly advanced stage, has its share of problems. We briefly mentioned some of them in the introduction, but it's time to define them in a more concise manner. Here we can enumerate the NeRFxSatellite problems as following:

1. The shadow problem
2. The angle problem
3. The transient objects problem
4. The multi-date problem
5. The data sparsity problem

In this section we will talk about how NeRFs relate to satellite imagery with respect to all their problems by analyzing the nature of NeRFs and the anatomy of satellite imagery, hereby focusing on what we would call the two fundamental

models of NeRFs for satellites in the literature, namely Shadow-NeRF (S-NeRF) and Sat-NeRF.

The original NeRF approach is destined for on-ground, localized computer science applications rather than off-nadir satellite angle reconstructions.

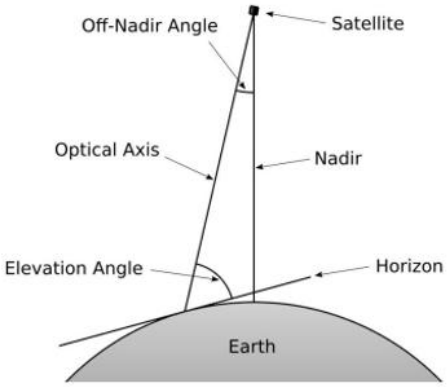


Fig. 2. Important angles for satellite imagery.

The nature of the WorldView-3 satellite dataset is such that the images exhibit strong non-correlation because they are taken at different moments in time. Different dates for photos come with different lighting conditions. This is how **multi-date problem** manifests itself when trying to apply vanilla NeRF. Shadows moving in between images can cause errors to manifest when using altitude estimation methods based on pixel or feature matching.

When trying to solve **the shadow problem**, it's hard to get off the ground - because shadows cannot be easily detected due to ambient illumination from the sky.

Another thing we observe from the benchmark dataset is the fact that the number and distribution of viewing angles is restricted. In other words, we have the **data sparsity** as well as the **angle problem**, as defined above. Because the dataset was collected with traditional stereo-matching methods in mind, WorldView-3 dataset contains an average of 10-20 images with and a sparse distribution of solar angles, with off-nadir angle lower than 35°.

Seen from the perspective of the original NeRF algorithm, the vanilla Neural Radiance Fields model is not built for satellite images, because of many reasons, spanning from the specificities of the camera models, the large distance between the cameras and the scene or the appearance inconsistencies of multi-date collections [24][25].

NeRF assumes that the density, radiance and illumination of the target 3D scene is constant. This is a strong limitation, as these conditions are rarely encountered outside laboratory settings. We will define the two fundamental models that are designed to resolve these issues.

B. Benchmarks: The two key models

FM1: Shadow-NeRF (S-NeRF)

The vanilla Neural Radiance Fields (NeRF) model simultaneously represents scene geometry and color using a volumetric function:

$$F(x, \theta, \phi) \rightarrow (c, \sigma)) \tag{1}$$

$$f(x) = (\chi_\lambda, \sigma) \tag{2}$$

where $x \in \mathbb{R}^3$ represents 3D spatial coordinates, c_λ denotes the outgoing radiance, and $\sigma \in \mathbb{R}^+$ represents volume density. Traditional NeRF does not explicitly model lighting conditions, leading to inaccurate shadow reconstructions. Shadow-NeRF (S-NeRF) introduces a spectral irradiance vector $\ell_\lambda(\mathbf{x}, \omega_s)$, which models incoming light at $\mathbf{x}$ for a given solar $\omega_s = (\theta_s, \varphi_s)$ [25]. Two additional network outputs are introduced: (1) $s(\mathbf{x}, \omega_s)$ – ratio of direct solar illumination to diffuse sky light, modeling shadow visibility at $\mathbf{x}$ where $s = 0$ represents complete shadow, while $s = 1$ denotes full sunlight exposure; (2) $\mathbf{sky}(\omega_s)$ – a learned vector representing the sky's diffuse illumination, independent of spatial location.

The total irradiance $\ell(\mathbf{x}, \omega_s)$ is computed as:

$$\ell = s1_3 + (1 - s)\mathbf{sky} \tag{3}$$

Instead of modeling radiance $c(x)$ directly, S-NeRF predicts albedo a(x) and computes radiance as:

$$\mathbf{c}(\mathbf{x}, \omega_s) = \mathbf{a}(\mathbf{x}) * \ell(\mathbf{x}, \omega_s) \tag{4}$$

This Lambertian reflectance model enhances shadow reconstruction by distinguishing albedo from illumination effects.

FM2: Sat-NeRF

Sat-NeRF extends NeRF to satellite imagery by incorporating solar illumination metadata and transient object handling. It takes as input:

- $\mathbf{x}$: 3D spatial coordinates.
- $\boldsymbol{\omega}$: Solar direction vector, extracted from azimuth and elevation metadata.
- $\boldsymbol{t}_j : N^{(t)}$: A learned embedding vector encoding transient elements unique to the j-th view.

The model outputs:

- σ: Volume density.
- $\mathbf{c}_a$: Albedo color, independent of illumination.
- s: Shadow-aware shading scalar, dependent on x and the solar rays direction vector $\boldsymbol{\omega}$.
- $\mathbf{a}$: Ambient RGB color, capturing global shading effects.
- β: Uncertainty coefficient for transient objects.

Sat-NeRF's radiance function is defined as:

$$c(x, \omega, t_j) = c_a(x) \cdot s(x, \omega) + (1 - s(x, \omega)) \cdot a(\omega)) \quad (5)$$

where $s(x, \omega)$ darkens the albedo in shadowed areas, and $a(\omega)$ captures ambient illumination effects.

To address solar angle dependencies, a **solar correction loss** L_{SC} is introduced:

$$L_{SC}(R_{SC}) = \sum_{r \in R_{SC}} (\sum_{i=1}^{N_{SC}} (T_i - s_i)^2 + 1 - \sum_{i=1}^{N_{SC}} T_i \alpha_i s_i) \quad (6)$$

where R_{SC} contains solar rays parallel to the sun's direction, ensuring $s(x, \omega)$ generalizes across lighting conditions.

To mitigate transient object artifacts, Sat-NeRF employs an **uncertainty weighting mechanism**, modeled by $\beta(\mathbf{x}_i, \mathbf{t}_j)$, which assigns lower weights to inconsistent color observations:

$$\beta(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \beta(\mathbf{x}_i, \mathbf{t}_j) \quad (7)$$

C. Datasets

Most of the satellite NeRF papers (SI1, SI2, SI3, SI4, SI9) rely on the 2019 IEEE GRSS Data Fusion Contest dataset, which consists of multi-view WorldView-3 imagery acquired between 2014 and 2016 over Jacksonville, Florida, USA. In contrast, models like SI6 (Planet-NeRF) and SI10 (Season-NeRF) extend this by including additional areas from Omaha, Nebraska, USA—thus introducing a component of seasonal variation. Other models incorporate entirely different datasets: SI5 (EO-NeRF) combines the DFC dataset (WorldView-3 over Jacksonville) with data from the 2016 IARPA Multi-View Stereo 3D Mapping Challenge covering the Buenos Aires area in Argentina, and SI7 (BRDF-NeRF) uses datasets from Djibouti (Asal-Ghoubbet rift) and Lanzhou (China). Finally, SI8 (Sparsesat-NeRF) leverages both the Djibouti dataset and DFC data from Jacksonville, while SI11 (Sat-Mesh) uses WorldView-3 images (with a 0.3m GSD) covering both Jacksonville and Omaha.

This variation in datasets is significant because it influences the appearance characteristics (seasonal variations, transient phenomena, and resolution) that each model must handle. For instance, models developed for Jacksonville (mostly urban areas) might emphasize shadow and transient object handling, whereas those incorporating Omaha data are more concerned with seasonal changes. Models using data from Djibouti and Lanzhou further address unique surface reflectance properties and challenging illumination conditions. We will only include a description of two out of the four datasets, for the sake of brevity.

2019 IEEE GRSS Data Fusion Contest dataset

Since the publication of the two pioneering papers on this subject, 2019 IEEE GRSS Data Fusion Contest dataset has served as the benchmark for all studies pertaining to applications of NERFs to satellite imagery 3D reconstructions.

The dataset first appeared as part of the Data Fusion Contest (DFC), organized by IEEE and Remote Sensing Society (GRSS). The main idea of the contest addressed 3D modeling and automated cartography. For these purposes, the data included multi-view, multiband satellite images; airborne lidar data; and semantic geospatial information about roads, bridges, buildings, trees, and water.

The contest made use of Urban Semantic 3D data, containing more than 320GB of data, covering approximately 20km^2 over the urban areas of Jacksonville, Florida, and Omaha, Nebraska [26]. Precisely, it comprised the following:

- Multidate satellite images: WorldView-3 panchromatic and eight-band visible and near-infrared images were provided (courtesy of DigitalGlobe), with ground-sampling distances of 35cm and 1.3m, respectively.
- Airborne lidar data: These are made publicly available by the U.S. Homeland Security Infrastructure Program. They were provided as point clouds or digital surface models (DSMs), with an aggregate nominal pulse spacing of approximately 80cm.
- Semantic labels: These labels (buildings, elevated roads and bridges, high vegetation, ground, and water) were given for each geographic tile, as either semantic maps or point labels for point clouds.

Djibouti dataset

This dataset is used by two of the mentioned papers, namely [S7 – BRDF-NeRF] and [S8 – Sparsesat-NeRF].

It represents a series of 21 multi-angular Pleiades images collected in a single flyby on January 26, 2013 located in the Asal-Ghoubbet rift (Republic of Djibouti). One of the reasons this site was chosen is precisely because of its significance - it displays a wide variety of bare soil over a small area (≈100km^2), from pure white salt deposits to dark lava basalt flows, a contrasted topography, and a great wealth of surfaces, from very rough lava flows to flat deposits in sedimentary basin. The quadcopter was maintained at a flight altitude between 8m and 15m to obtain a pixel size on the ground ranging from 1 to 2mm. In the SI8 paper [14], for the training purposes only, two/three RGB cropped images (≈800×800px), with 2m Ground Sampling Distance (GSD) were used. Whereas in the SI7 paper [16], two regions within the rift are chosen, and from them - three quasi-nadir images are chosen for training, and three other images for testing, including interpolation and extrapolation scenarios.

IV. Results & Discussion

A. Categorization by problems addressed

Inspired by the five problems we identified in section III, we've decided to come up with the following preliminary categorization of the NeRFxSatellite models. We group NeRFxSatellite models (represented by the {SI} papers) in the following way:

1) Shadow-Aware and Transient-Robust NeRFs for Satellite Imagery (addresses Problem1 & Problem2)

Basic Shadow Modeling:

SI1 – Surf-NeRF: Adapts NeRF to handle off-nadir satellite images by explicitly modeling the lighting (sun and sky) to "explain" shadows [27].

SI3 – Shadow NeRF (S-NeRF): The first effort to incorporate solar visibility into a NeRF framework for multi-view satellite photogrammetry.

Enhanced with Camera Models and Transients:

SI4 – Sat-NeRF: Builds on S-NeRF by using native RPC (Rational Polynomial Coefficient) camera models and by adding components to handle transient objects (like cars or vegetation changes).

SI5 – EO-NeRF: Further refines the approach by enforcing geometric consistency in shadow rendering, thereby boosting DSM accuracy.

2) Extensions to Capture Seasonal and Reflectance Variations (addresses Problem3 and Problem4)

Seasonal Variation:

SI6 – Planet-NeRF: Extends Sat-NeRF by incorporating a set of month-embedding vectors so that the network can learn to render seasonal changes (e.g., snow versus vegetation).

SI10 – Season-NeRF: Also adds time-of-year as an input to disentangle seasonal appearance from solar shadow effects, ensuring that seasonal features (like foliage or snow) change only with time [29].

Advanced Reflectance Modeling:

SI7 – BRDF-NeRF: Integrates an explicit physically-based BRDF model (the Rahman-Pinty-Verstraete (RPV) model) into NeRF to better capture the anisotropic reflectance properties of natural surfaces.

3) Efficiency and Sparse-View Adaptations (addresses mostly Problem5)

Speed & Efficiency:

SI2 – SAT-NGP: Uses ideas from Instant-NGP (like multi-resolution hash encoding) to speed up training dramatically (reducing training time from hours to minutes) while still performing 3D reconstruction [28].

Sparse View and Dense Depth Supervision:

SI8 – Sparsesat-NeRF: Adapts Sat-NeRF for scenarios with very few satellite views by incorporating dense depth supervision (using SGM-derived depth maps) and guided ray sampling.

SI9 – SatelliteRF: Focuses on accelerating the overall 3D reconstruction process using multi-resolution hash encoding and a smaller MLP, thereby cutting training time from 8–10 hours to roughly 5–10 minutes.

4) Direct Mesh Generation / Implicit Surface Reconstruction (addresses all problems, including Problem5)

Direct Mesh Extraction:

SI11 – Sat-Mesh: Instead of producing a 2.5D DSM (as in most NeRF variants), Sat-Mesh learns an implicit surface (a signed distance function, or SDF) directly from the images using a volume-rendering framework. The zero-level set is then extracted (via marching cubes) to form a full 3D mesh. It also introduces a multi-view stereo (MVS) constraint (using normalized cross-correlation) and latent appearance encoding to capture seasonal variations for texturing [30].

This categorization is vital for the literature review as it provides a structured framework to systematically compare and contrast the various approaches developed for satellite imagery reconstruction. By clearly delineating how each model addresses the five problems we defined (See Section III) – from shadow and transient effects to seasonal variations, efficiency improvements, and direct mesh generation – it highlights the evolution of techniques and identifies the strengths and limitations inherent in each approach. This organization not only enhances understanding of current methodologies but also helps pinpoint areas for future research and development.

B. Categorization by selected datasets

TABLE II. NERFxSATELLITE MODELS AND THEIR DATASETS

Model (SI#)	Datasets Used
SI1 – Surf-NeRF	2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016)
SI2 – SAT-NGP	2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016)
SI3 – S-NeRF	2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016)
SI4 – Sat-NeRF	2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016)
SI5 – EO-NeRF	2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016); 2016 IARPA MVS 3D Mapping Challenge
SI6 – Planet-NeRF	2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016);
SI7 – BRDF-NeRF	Djibouti dataset; Lanzhou dataset
SI8 – Sparsesat-NeRF	Djibouti dataset; 2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016)
SI9 – SatelliteRF	2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016)
SI10 – Season-NeRF	2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016)
SI11 – Sat-Mesh	2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016) with 0.3m GSD

Another categorization of models included in this study is given in TABLE III. This categorization is based on the datasets used by each model. It is important to note that 2019 IEEE GRSS Data Fusion Contest, WorldView-3 (2014–2016) is to be found in almost all the datasets, with the photos taken from the geographic region of Jacksonville, FL, USA, with the following exceptions: SI6 and SI11 use this dataset for both Jacksonville, FL, USA and Omaha, NE, USA, whereas SI10 excludes Jacksonville and includes only the images of the geographic region Omaha, NE, USA for the same dataset. This categorization is significant because it facilitates further studies so as to allow a comparative analysis based on quantifiable metrics of all models mentioned in this study.

V. Conclusion & Future Work

In this paper, we examined the application of Neural Radiance Fields (NeRFs) to satellite imagery, analyzing a dozen papers – each introducing its own model. We focused on two fundamental models and explored how they modified the vanilla NeRF framework to meet the specific challenges of satellite imaging. Additionally, we identified five key issues inherent in NeRFxSatellite models and provided a categorization that illustrates how each approach attempts to address these challenges.

This structured analysis not only enhances our understanding of current methodologies but also lays the groundwork for future research. Although a standardized, quantitative comparative analysis – using metrics such as MAE (Mean Absolute Error), MED (Median Error), and Perc-1m (Percentage within 1 meter) – could further assess the efficacy of each model, such an evaluation remains beyond the scope of this study. Nonetheless, this paper provides a clear roadmap for subsequent studies aimed at refining and benchmarking these innovative techniques in satellite imagery processing.

References

[1] NASA. (n.d.). The problem of Space Travel: The rocket motor - NASA technical reports server (NTRS). NASA. https://ntrs.nasa.gov/citations/19950015477A [The Problem of Space Travel, pg.119]

[2] Andy and Andy, "How many earth observation satellites are orbiting the planet in 2021?," Pixalytics Ltd, https://www.pixalytics.com/eo-sats-2021/ (accessed March.3 1, 2025).

[3] Huang, X.; Wen, D.; Li, J.; Qin, R. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. *Remote Sens. Environ.* **2017,** *196*, 56–75.

[4] Li, S.; Zhu, Z.; Wang, H.; Xu, F. 3D virtual urban scene reconstruction from a single optical remote sensing image. *IEEE Access* **2019,** *7,* 68305–68315.

[5] Zhao, Q., Yu, L., Du, Z., Peng, D., Hao, P., Zhang, Y., & Gong, P. (2022). An Overview of the Applications of Earth Observation Satellite Data: Impacts and Future Trends. *Remote Sensing*, 14(8), 1863. https://doi.org/10.3390/rs14081863

[6] Huang, Y.; Dugmag, H.; Barfoot, T.D.; Shkurti, F. Stochastic planning for asv navigation using satellite images. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; IEEE: New York, NY, USA, 2023; pp. 1055–1061

[7] Smith, W., Bishop, W., Dvorak, V., Hayden, C., McElroy, J., Mosher, F., Oliver, V., Purdom, J., Wark, D.: The Meteorological Satellite: Overview of 25 Years of Operation. Science. (1986)

[8] Barrile, V.; Bilotta, G.; Fotia, A.; Bernardo, E. Road extraction for emergencies from satellite imagery. In *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, 1–4 July 2020, Proceedings;* Part IV 20; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 767–781.

[9] Liu, C.; Szirányi, T. Road Condition Detection and Emergency Rescue Recognition Using On-Board UAV in the Wildness. *Remote Sens.* **2022,** *14,* 4355

[10] Dakir, A., Barramou, F., Alami, O.B.: Opportunities for Artificial Intelligence in Precision Agriculture Using Satellite Remote Sensing. Geospatial Intelligence: Applications and Future Trends pp. 107–117 (2022)

[11] A Ravanbakhsh, M., Ramzi, P.: Creating Forest Inventory from HighResolution Satellite Images. https://blog.maxar.com/earth-intelligence/2018/creating-forest-inventory-from-high-resolution-satellite-images (12032018)

[12] Höschle, C., Cubaynes, H.C., Clarke, P.J., Humphries, G., Borowicz, A.: The Potential of Satellite Imagery for Surveying Whales. Sensors **21**(3) (2021). https://doi.org/10.3390/s21030963, https://www.mdpi.com/1424-8220/21/3/963

[13] El-Gamal F.E.-Z.A., Elmogy M., Atwan A. Current Trends in Medical Image Registration and Fusion. Egypt. Inform. J. 2016;17:99–124. doi: 10.1016/j.eij.2015.09.002.

[14] Lulin Zhang, Ewelina Rupnik. SparseSat-NeRF: Dense Depth Supervised Neural Radiance Fields for Sparse Satellite Images. ISPRS Annals 2023, 2023. ffhal-04205387f [SI8]

[15] Zhou, X., Wang, Y., Lin, D., Cao, Z., Li, B., & Liu, J. (2024). SatelliteRF: Accelerating 3D Reconstruction in Multi-View Satellite Images with Efficient Neural Radiance Fields. Applied Sciences, 14(7), 2729. https://doi.org/10.3390/app14072729 [SI9]

[16] L. Zhang, E. Rupnik, D. Nguyen, S. Jacquemoud, Y. Klinger. (2024). BRDF-NeRF: Neural Radiance Fields with Optical Satellite Images and BRDF Modelling. 10.48550/arXiv.2409.12014. [SI7]

[17] R. Marí, G. Facciolo and T. Ehret, "Multi-Date Earth Observation NeRF: The Detail Is in the Shadows," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023, pp. 2035-2045, doi: 10.1109/CVPRW59228.2023.00197. [SI5]

[18] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In Computer Graphics Forum, volume 39, pages 701–727, 2020. 1

[19] *Digital Surface Model (DSM)*. Datumate. (2024, August 29). https://www.datumate.com/glossary/digital-surface-model-dsm/

[20] Fuentes Reyes, M.; d'Angelo, P.; Fraundorfer, F. Comparative Analysis of Deep Learning-Based Stereo Matching and Multi-View Stereo for Urban DSM Generation. *Remote Sens*. 2025, 17, 1. https://doi.org/10.3390/rs17010001

[21] Hu, Z., Tao, P., Long, X., & Wang, H. (2022). Shading aware DSM generation from high resolution multi-view satellite images. *Geo-Spatial Information Science*, *27*(2), 398–407. https://doi.org/10.1080/10095020.2022.2125837

[22] Yin, L.; Wang, L.; Zheng, W.; Ge, L.; Tian, J.; Liu, Y.; Yang, B.; Liu, S. Evaluation of empirical atmospheric models using Swarm-C satellite data. *Atmosphere* **2022**, *13*, 294.

[23] Kåreborn, Liv & Ingerstad, Erica & Berg, Amanda & Karlsson, Justus & Haglund, Leif. (2024). Exploring Seasonal Variability in the Context of Neural Radiance Fields for 3D Reconstruction on Satellite Imagery. 10.48550/arXiv.2411.02972.[SI6]

[24] Marí, Roger & Facciolo, Gabriele & Ehret, Thibaud. (2022). Sat-NeRF: Learning Multi-View Satellite Photogrammetry With Transient Objects and Shadow Modeling Using RPC Cameras. 1310-1320. 10.1109/CVPRW56347.2022.00137. [SI4]

[25] Derksen, Dawa & Izzo, Dario. (2021). Shadow Neural Radiance Fields for Multi-view Satellite Photogrammetry. 1152-1161. 10.1109/CVPRW53098.2021.00126. [SI3]

[26] M. Bosch, G. Foster, G. Christie, S. Wang, G. D. Hager, M. Brown, "Semantic stereo for incidental satellite images," in *Proc. Winter Conf. Applications of Computer Vision*, 2019.

[27] Semeraro, F., Zhang, Y., Wu, W., & Carroll, P. (2023, April 18). Nerf applied to satellite imagery for surface reconstruction. arXiv.org. https://arxiv.org/abs/2304.04133 [SI1]

[28] Billouard, C., Derksen, D., Sarrazin, E., & Vallet, B. (2024, March 27). Sat-NGP : Unleashing neural graphics primitives for fast relightable transient-free 3D reconstruction from satellite imagery. arXiv.org. https://arxiv.org/abs/2403.18711 [SI2]

[29] Billouard, C., Derksen, D., Sarrazin, E., & Vallet, B. (2024, March 27). Sat-NGP : Unleashing neural graphics primitives for fast relightable transient-free 3D reconstruction from satellite imagery. arXiv.org. https://arxiv.org/abs/2403.18711 [SI10]

[30] Qu, Y., & Deng, F. (2023). Sat-Mesh: Learning Neural Implicit Surfaces for Multi-View Satellite Reconstruction. *Remote Sensing, 15*(17), 4297. https://doi.org/10.3390/rs15174297 [SI11]

Software Implementations of Lightweight Ciphers: A Survey with Focus on ASCON

1st Ikechukwu John Chukwu
Faculty of Computer Science and Engineering
Ss. Cyril & Methodius University
Skopje, North Macedonia
ikechukwu.chukwu@students.finki.ukim.mk

2nd Vesna Dimitrova
Faculty of Computer Science and Engineering
Ss. Cyril & Methodius University
Skopje, North Macedonia
vesna.dimitrova@finki.ukim.mk

***Abstract*—Lightweight cryptography is indispensable for securing pervasive devices with limited resources, such as those found in the Internet of Things (IoT), RFID, healthcare, and industrial applications. The proliferation of these technologies necessitates robust yet efficient security solutions, often implemented in software within resource-constrained architectures.**

This paper presents a survey focusing on *software implementations* of lightweight ciphers. We trace the evolution of these ciphers, categorizing them into pre-NIST and post-NIST (National Institute of Standards and Technology Lightweight Cryptography competition) eras. We review ASCON's design principles and discuss notable software implementations and security considerations.

***Index Terms*—Cryptography, Lightweight ciphers, ASCON**

I. INTRODUCTION

Despite their "lightweight" nature, birds soar to great heights. Similarly, in the secure communication domain, while robust, sophisticated security layers are often the expectation, the proliferation of resource-constrained devices has brought "lightweight" cryptographic solutions to the forefront. This naturally raises the question: Does focusing on "lightweight" imply a compromise in security?

The answer is a firm no. In cryptography, "lightweight" refers to security solutions meticulously optimized for pervasive devices. These devices operate with stringent limitations on energy for computation, memory for storage, and processing power, yet they still demand strong, unexploitable security.

Security primitives provide crucial functions in computation, preventing malicious actors from exploiting system vulnerabilities. Designing such primitives requires significant theoretical and analytical effort. However, a one-size-fits-all security design is often impractical; solutions may excel in hardware but perform poorly in software, or vice versa. This dichotomy has long been a concern for both hardware and software communities, especially with the rise of pervasive devices.

To address this need, the academic and research community has extensively reviewed existing ciphers and explored advancements in lightweight cryptography. Some efforts have surveyed ciphers deemed lightweight [1]–[5], while others have focused on implementing and benchmarking specific ciphers on various architectures to assess their performance and feasibility [6]–[8].

A pivotal moment was the National Institute of Standards and Technology (NIST) Lightweight Cryptography (LWC) competition. Competition for Authenticated Encryption: Security, Adaptability, and Robustness (CEASAR) was initiated with a call for submissions in 2018 and officially launched in 2019, following their 2015 survey and 2017 report [2]. The competition aimed to standardize one or more lightweight authenticated encryption with associated data (AEAD) schemes and, optionally, hashing schemes. After multiple evaluation rounds, the ASCON family was announced as the winner in February 2023 [9].

ASCON was selected for its strong security, excellent performance across both hardware and software platforms, and overall suitability for constrained environments. This paper surveys software implementations of lightweight ciphers, with a specific focus on ASCON. Our aim is to provide insights into its performance characteristics by contextualizing it within the broader landscape of pre-NIST and contemporary lightweight ciphers. We will touch upon key software performance metrics such as throughput, latency, code size, and RAM usage, which are critical for resource-constrained environments like IoT devices and embedded systems.

The remainder of this paper is organized as follows. Section II reviews fundamental terms and techniques, providing essential terminology for cryptography and lightweight ciphers. Section III discusses software implementations of lightweight ciphers, covering their evolution and key findings. Finally, Section IV concludes the paper and outlines future research directions.

II. BACKGROUND

In this section, we introduce the fundamental concepts necessary to understand the subsequent discussion on lightweight ciphers and their software implementations.

A. Cryptography

The primary objective of cryptography is to hide data and prevent all forms of malicious access, interference, and recovery of data. As man increases his empire, more security needs arise. The improvements in cryptography have seen improvements in exploitation. For example, the case of public-key cryptosystems and quantum computation. However,

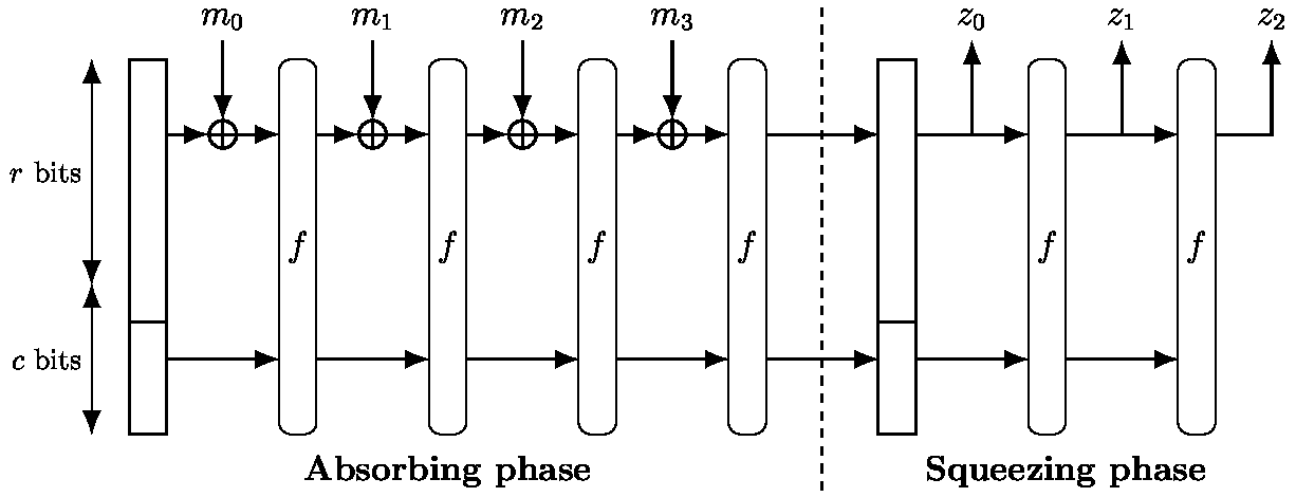


Fig. 1. Sponge structure (adapted from IACR TikZ collection)

cryptography still remains a fascinating domain with a lot of untapped options.

Modern cryptography includes symmetric-key algorithms (using the same key for encryption and decryption, like block ciphers and stream ciphers), asymmetric-key algorithms (using public/private key pairs, like RSA and ECC), and cryptographic protocols (combining primitives to achieve complex security goals, like TLS/SSL). This survey primarily focuses on symmetric-key block ciphers designed for lightweight applications.

B. Lightweight Ciphers and Software Metrics

Lightweight cryptography refers to cryptographic primitives specifically designed and optimized for resource-constrained environments, such as Internet of Things (IoT) devices, RFID tags, embedded systems, and sensor nodes [2]. These devices typically operate with significant limitations on processing power, memory (both program and operational), energy consumption, and physical size.

Lightweight ciphers can be implemented in dedicated hardware circuits or as software running on general-purpose microcontrollers. While hardware implementations often prioritize minimizing gate count, area, and power [2], software implementations face a different set of critical constraints. This survey focuses on software implementations, for which the following performance metrics are particularly important:

1) Throughput: Throughput measures the rate at which data can be processed (encrypted or decrypted) by the cipher. It is a crucial metric for applications that handle streams of data or require high processing speeds. In software benchmarks, it is often reported in cycles per byte (c/B), where a lower value indicates higher efficiency, as it represents the number of processor clock cycles needed to process each byte of data. Alternatively, it can be expressed in kilobits per second (kbps) or Megabits per second (Mbps) for a given processor clock frequency. Higher throughput generally means better performance for bulk data operations.

2) Latency: Latency refers to the time delay, measured in clock cycles, required to complete a single cryptographic operation, such as encrypting one block of plaintext or performing an authentication tag generation. Low latency is essential for applications requiring real-time responsiveness or involving frequent operations on small, individual data packets. Unlike throughput, which measures bulk processing speed, latency measures the delay for a discrete operation.

3) Code Size: Code size represents the amount of non-volatile memory, typically Read-Only Memory (ROM), needed to store the compiled instructions of the cryptographic algorithm. This is measured in bytes or Kilobytes (KB). Minimizing code size is critical for microcontrollers with limited program memory, as it allows more space for other application logic or enables the use of cheaper devices with smaller memory footprints.

4) RAM Usage: Random Access Memory (RAM) usage indicates the amount of volatile memory required by the algorithm during execution. This includes memory for storing the internal cipher state, round keys (key schedule), temporary variables, and function call stack space. It is measured in bytes or Kilobytes (KB). Low RAM usage is often a stringent requirement, as RAM is typically the most scarce memory resource on low-end embedded devices. Excessive RAM usage can make a cipher infeasible to implement on certain platforms.

C. ASCON

The ASCON family of algorithms was selected as the winner of the NIST Lightweight Cryptography (LWC) competition in February 2023, representing a significant development for securing resource-constrained devices [10]. Its selection was based on a combination of strong security, excellent perfor-

mance across various platforms (both hardware and software), and flexibility.

A core design principle of the ASCON family is its reliance on a single, lightweight cryptographic permutation, denoted as p. This permutation operates on a 320-bit internal state and is constructed using only Substitution-Permutation Network (SPN) principles with efficient bitwise operations (rotation, XOR). Crucially, the permutation is inverse-free, meaning the inverse operation is not needed for decryption or hashing. This significantly simplifies implementation, reducing code size in software and gate count in hardware compared to designs requiring distinct forward and inverse functions (like AES).

ASCON's AEAD and hashing modes are built upon the Sponge construction [11]. A generic Sponge construction is illustrated in Figure 1. In this model, the 320-bit state is divided into an outer part, the rate (r), and an inner part, the capacity (c). The capacity (c) provides security, while the rate (r) determines the block size for processing data. For ASCON-128, $r = 64$ bits and $c = 256$ bits.

The ASCON-AEAD128 authenticated encryption process proceeds in phases:

1) **Initialization:** The 128-bit key (K) and the 128-bit nonce (N) are absorbed into the initial state. The permutation p is applied $a = 12$ times.
2) **Processing Associated Data (AD):** The associated data (if any) is absorbed into the state in r-bit blocks. After absorbing each block of AD, the permutation p is applied $b = 6$ times.
3) **Processing Plaintext:** The plaintext is absorbed into the state in r-bit blocks. Simultaneously, the corresponding r-bit ciphertext blocks are squeezed from the state. After processing each block, the permutation p is applied $b = 6$ times.
4) **Finalization:** The key is absorbed into the state again. The permutation p is applied $a = 12$ times. Finally, a 128-bit authentication tag is squeezed from the state.

Decryption follows a similar process, re-computing the tag and comparing it to the received tag to ensure both confidentiality and integrity.

ASCON demonstrated robust security throughout the extensive public scrutiny of the NIST competition, withstanding numerous cryptanalysis attempts [9]. This resilience, combined with its efficiency, makes it a highly suitable candidate for widespread adoption in diverse applications, from IoT devices and embedded systems to potentially even broader use cases.

The following section will delve into software implementations, discussing pre-NIST ciphers and providing context for evaluating ASCON's performance relative to its predecessors and contemporaries.

III. Software Implementations of Lightweight Ciphers

This section reviews software implementations of lightweight ciphers, tracing some developments preceding the official NIST Lightweight Cryptography (LWC) competition, and implementations emerging after ASCON was announced as the winner. We will examine these implementations primarily through the lens of software performance metrics relevant to resource-constrained devices, such as code size, RAM usage, throughput, and latency, as discussed in Section II-B.

A. Pre-NIST Implementations and Benchmarking Efforts

Cryptographic algorithms are fundamental to securing data on diverse software platforms. In the context of lightweight cryptography, while many devices are hardware-based, they often incorporate software components or are emulated in software environments. Consequently, understanding software performance is crucial.

Several notable studies have investigated and benchmarked software implementations of lightweight ciphers on constrained platforms. For instance, Sevin and Mohammed [8] conducted a comprehensive survey involving the implementation of 50 lightweight block ciphers, all predating the NIST LWC call, on an 8-bit AVR microcontroller (simulated using Atmel Studio 7). Their analysis on an emulated ATmega128 (4KB RAM, 128KB Flash) focused on code size and RAM usage as primary performance indicators. Their findings indicated that CHASKEY [38], an ARX-based (Addition-Rotation-XOR) cipher designed for 32-bit microcontrollers, exhibited favorable speed performance in this 8-bit simulated environment.

Another significant benchmarking effort by Raza et al. [39] focused on the block cipher LED [32], which shares structural similarities with AES. They explored software implementations in Java, C#, C++, and Python, deploying them on Amazon EC2 cloud instances (32-bit and 64-bit Linux/Windows). This study highlighted the performance variations across different programming languages and operating environments, even for a cipher like LED, which, while accommodating software, was initially conceived with hardware considerations.

A broader review by Hatzivasilis et al. [3] summarized software implementation characteristics of numerous pre-NIST lightweight ciphers, many of which are also listed in our overview Table I. This review highlighted that ciphers like SPECK, SIMON, AES, PRIDE, SEA, and XTEA achieved particularly compact implementations in terms of code size. Furthermore, considering a wider range of metrics including ROM usage, RAM usage, latency (cycles), energy consumption (μJ), and throughput (kbps), they identified strong performers for different microcontroller architectures:

- On 8-bit microcontrollers: DESXL-184, SPECK-128, PRIDE-128, AES-128, Fantomas-128, PRESENT-128, ROBIN-128, SPECK-96, SEA-96, ITUbee-80, PRESENT-80, and DESL-56 were noted for efficiency.
- On 16-bit devices: DESXL-184, Fantomas-128, Robin-128, AES-128, PRINCE-128, SPECK-96, SEA-96, and PRESENT-80 performed well.
- On 32-bit microcontrollers: Fantomas-128, Robin-128, AES-128, SPECK-96, and PRESENT-80 were considered among the best.

TABLE I
OVERVIEW OF SELECTED LIGHTWEIGHT CIPHERS BY STRUCTURE (FEISTEL, SPN, SPONGE) [8]

Cipher	Block Size (bits)	Key Size (bits)	Rounds	Structure	Primary Target[a] (SW/HW)	Year	Reference(s)
Feistel Network Based							
RC5	32/64/128	0–2040	1–255	Feistel	SW	1994	[12]
XTEA	64	128	64 (32 cycles)	Feistel	SW	1997	[13]
SKIPJACK/KEA	64	80	32	Feistel	SW/HW	1998	[14], [15]
SEED	128	128	16	Feistel	SW	1998	[16]
RC6	128	128/196/256	20	Feistel	SW	1998	[17], [18]
SEA	48/96/144	48/96/144	51/93/135	Feistel	SW	2006	[19]
PRESENT	64	80/128	31	Feistel	HW	2007	[20]
TWINE	64	80/128	36	Feistel	SW/HW	2011	[21]
PICCOLO	64	80/128	25/31	Feistel	HW	2011	[22]
SM4	128	128	32	Feistel	HW	2012	[23]
SPECK[b]	32–128	64–256	22–34	Feistel	SW/HW	2013	[24]
SIMON[b]	32–128	64–256	32–72	Feistel	SW/HW	2013	[24]
ROADRUNNER-R	64	80/128	10/12	Feistel	SW	2015	[25]
SPARX	64/128	128/256	24/32/40	Feistel	SW	2016	[26]
Substitution-Permutation Network (SPN) Based							
SERPENT	128	128/196/256	32	SPN	HW	1998	[27]
SAFER++	128	128/196/256	8/12/16	SPN	SW	1998	[28]
AES	128	128/196/256	10/12/14	SPN	SW/HW	1998	[29]
mCrypton	64	64/96/128	12 (+1)	SPN	HW	2006	[30]
THREEFISH	256/512/1024	256/512/1024	72/80	SPN	SW/HW	2008	[31]
LED	64	64/128	32/48 (steps)	SPN	SW/HW	2011	[32]
PRINCE	64	128	12	SPN	SW	2012	[33]
ZORRO	128	128	24	SPN	SW	2013	[34]
PRIDE	64	128	19 (+1)	SPN	SW	2014	[35]
RECTANGLE	64	80/128	25	SPN	SW/HW	2015	[36]
Sponge Construction Based							
ASCON-AEAD128	64 (rate)	128	12/6/12	Sponge	SW/HW	2023	[37]

[a] Primary Target: Indicates if the cipher was primarily designed or performs best in Software (SW) or Hardware (HW), or both (SW/HW).
[b] SPECK and SIMON are families of block ciphers with varying block/key sizes. SPECK is generally more optimized for software performance, while SIMON often excels in hardware footprint [24].
[c] ASCON is a family of authenticated ciphers and hash functions. ASCON-AEAD128 is shown as the primary AEAD variant. Its internal state is 320 bits.
[d] ASCON-AEAD128 rounds: 12 for initialization (p_a), 6 for processing data blocks (p_b), 12 for finalization (p_a).

These pre-NIST studies established important benchmarks and demonstrated the trade-offs inherent in optimizing ciphers for software on constrained devices, paving the way for the requirements set forth in the NIST LWC competition.

B. ASCON Software Implementations and Post-NIST Context

The landscape of lightweight cryptography was significantly shaped by the NIST LWC competition, culminating in the selection of the ASCON family in February 2023. Even before its official selection, ASCON's strong performance during the competition rounds spurred early adoption and implementation studies.

An initial real-time deployment of ASCON was reported by the United States Air Force Academy (USAFA) in 2022 [40]. This implementation utilized Raspberry Pi Zero devices in a testbed featuring a weather messaging system and a smart-badge system, with communication over MQTT, a standard IoT messaging protocol. Performance analysis in this specific setup indicated that ASCON performed comparably to, albeit slightly less optimally than, AES-128 in terms of raw speed, but crucially, its encryption and decryption functionalities operated as expected, demonstrating its practical viability.

More recent work by Gigerl et al. [41] has focused on secure software implementations of ASCON, specifically addressing side-channel attack resilience. They developed **masked implementations of ASCON-AEAD128 for ARM and RISC-V microprocessors. Masking is a common countermeasure against power analysis attacks (which attempt to extract secret keys by observing a device's power consumption patterns) by splitting sensitive computations into multiple shares. Their work extended a 2-share design for the ASCON 5-bit S-box (χ) [42] to a 3-share version, demonstrating that robust masking for ASCON can be achieved efficiently, even without requiring additional sources of randomness for certain parts of the S-box computation, due to how random variables (r_0, r_1) are incorporated. This research underscores the ongoing efforts to provide not just efficient, but also secure software implementations of ASCON for practical deployment.

These examples illustrate the transition from general benchmarking of diverse lightweight ciphers to focused efforts on optimizing and securing the new NIST standard, ASCON, for real-world software applications.

IV. Discussion and Conclusion

This survey has provided a focused overview of software implementations of lightweight ciphers, with a particular emphasis on ASCON, the winner of the NIST Lightweight Cryptography (LWC) competition. We underscored the critical role of such cryptographic primitives in securing resource-constrained pervasive environments, prevalent in sectors like IoT, healthcare, and even military applications. Our review traced the evolution of lightweight cipher implementations, categorizing them into pre-NIST and post-NIST eras, an approach we believe offers valuable context for both academic researchers and industry practitioners.

We introduced ASCON's core design principles, including its reliance on an inverse-free permutation and sponge construction, which contribute to its efficiency and suitability for constrained devices. We then explored existing software implementation efforts, highlighting studies that reported on key efficiency metrics such as code size, RAM usage, and performance, and discussed initial deployments and security-enhancing implementations (like masking) of ASCON.

While a comprehensive, multi-cipher comparative benchmark table for software performance was beyond the scope of this particular survey due to its breadth, the presented information and cited works collectively point towards ASCON's strong potential in software. The studies on pre-NIST ciphers established baselines, and the initial ASCON implementations, along with NIST's own extensive evaluation, confirm its favorable characteristics for software deployment.

Looking ahead, several important avenues for future research emerge:

- **Comprehensive Benchmarking:** A crucial next step is the systematic benchmarking of ASCON's software performance (throughput, latency, code size, RAM usage) against other relevant lightweight ciphers (including NIST LWC finalists and established pre-NIST ciphers like SPECK) across a diverse range of common microcontroller architectures (e.g., 8-bit AVR, ARM Cortex-M series, RISC-V). Utilizing standardized frameworks like PlatformIO or FELICS for such evaluations would ensure comparability and provide invaluable data for developers.
- **Advanced Side-Channel Countermeasures:** Further exploration and optimization of side-channel attack countermeasures for ASCON in software, beyond the initial masking work cited [41], are essential. This includes investigating different masking schemes, threshold implementations, and countermeasures against fault attacks for various ASCON family members.
- **Cryptanalysis:** Continued third-party cryptanalysis of the ASCON permutation and its modes of operation remains vital to ensure its long-term security as it sees wider adoption.
- **Real-world Protocol Integration:** Investigating the performance and overhead of integrating ASCON into lightweight communication protocols (e.g., CoAP, MQTT-SN, LoRaWAN) and specific IoT application scenarios would provide practical insights into its real-world efficacy.

In conclusion, this survey offers valuable insights into the current landscape of lightweight cipher software implementations, with ASCON positioned as a highly promising standard. Future work, particularly in comprehensive benchmarking and robust security implementations, will be key to its successful and widespread adoption.

Acknowledgement

This work was partially supported by the European Union through the ERASMUS MUNDUS framework under the CyberMACS project (Project #101082683, https://cybermacs.eu/).

References

[1] B. J. Mohd, T. Hayajneh, and A. V. Vasilakos, "A survey on lightweight block ciphers for low-resource devices: Comparative study and open issues," *Journal of Network and Computer Applications*, vol. 58, pp. 73–93, Dec. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804515002076

[2] K. McKay, L. E. Bassham, M. S. Turan, and N. Mouha, "Report on Lightweight Cryptography," *NIST*, Mar. 2017. [Online]. Available: https://www.nist.gov/publications/report-lightweight-cryptography

[3] G. Hatzivasilis, K. Fysarakis, I. Papaefstathiou, and C. Manifavas, "A review of lightweight block ciphers," *Journal of Cryptographic Engineering*, vol. 8, no. 2, pp. 141–184, Jun. 2018. [Online]. Available: https://doi.org/10.1007/s13389-017-0160-y

[4] G. Pereira, M. Santos, B. Oliveira, M. Simplicio, P. Barreto, C. Margi, and W. Ruggiero, "SMSCrypto: A Lightweight Cryptographic Framework for Secure SMS Transmission," *Journal of Systems and Software*, Nov. 2012.

[5] A. Biryukov and L. Perrin, "State of the Art in Lightweight Symmetric Cryptography," 2017, publication info: Preprint. MINOR revision. [Online]. Available: https://eprint.iacr.org/2017/511

[6] K. Nekkanti, L. S. Sumanth, A. Renuka, M. Supriya, and A. Kumar, "Performance Analysis of Lightweight Ciphers for securing images," Dec. 2023, pp. 560–564. [Online]. Available: https://ieeexplore.ieee.org/document/10458747

[7] U. Panahi, C. Bayilmiş, çavuşoğlu, and S. Kacar, "Performance Evaluation of Lightweight Encryption Algorithms for IoT-Based Applications," *ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING*, Feb. 2021.

[8] A. Sevin and A. A. O. Mohammed, "A survey on software implementation of lightweight block ciphers for iot devices," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 3, pp. 1801–1815, 2023.

[9] C. Dobraunig, M. Eichlseder, F. Mendel, and M. Schläffer, "Ascon v1.2: Lightweight Authenticated Encryption and Hashing," vol. 34, no. 3, p. 33, Jun. 2021. [Online]. Available: https://doi.org/10.1007/s00145-021-09398-9

[10] T. Meltem Sönmez, M. Kerry A., C. Donghoon, K. Jinkeon, and K. John, "Ascon-Based Lightweight Cryptography Standards for Constrained Devices," *NIST Special Publication (SP)*, vol. 800, no. 232, 2024. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-232.ipd

[11] G. Bertoni, J. Daemen, M. Peeters, and G. V. Assche, "Cryptographic sponge functions," STMicroelectronics, Tech. Rep., 2007. [Online]. Available: http://sponge.noekeon.org/

[12] R. L. Rivest, "The rc5 encryption algorithm," in *Fast Software Encryption: Second International Workshop. Leuven, Belgium, 14-16 December 1994, Proceedings*, ser. Lecture Notes in Computer Science, vol. 1008. Springer, 1994, pp. 86–96.

[13] J. Chen, M. Wang, and B. Preneel, "Impossible differential cryptanalysis of the lightweight block ciphers tea, xtea and hight," in *Progress in Cryptology-AFRICACRYPT 2012: 5th International Conference on Cryptology in Africa, Ifrance, Morocco, July 10-12, 2012. Proceedings 5*. Springer, 2012, pp. 117–137.

[14] NIST, "Skipjack and kea algorithm specifications," *NIST*, 1998. [Online]. Available: https://csrc.nist.gov/csrc/media/projects/cryptographic-algorithm-validation-program/documents/skipjack/skipjack.pdf

[15] L. Granboulan, "Flaws in differential cryptanalysis of skipjack," Cryptology ePrint Archive, Paper 2001/038, 2001. [Online]. Available: https://eprint.iacr.org/2001/038

[16] H. Lee, S. Lee, J. Yoon, D. Cheon, and J. Lee, "Rfc 4269: The seed encryption algorithm," 2005.

[17] T. Shimoyama, M. Takenaka, and T. Koshiba, "Multiple linear cryptanalysis of a reduced round rc6," in *International Workshop on Fast Software Encryption*. Springer, 2002, pp. 76–88.

[18] R. L. Rivest, M. J. Robshaw, R. Sidney, and Y. L. Yin, "The rc6tm block cipher," in *First advanced encryption standard (AES) conference*, 1998, p. 16.

[19] M. Fouquet, P. Gaudry, and R. Harley, "Finding secure curves with the satoh-fgh algorithm and an early-abort strategy," in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2001, pp. 14–29.

[20] B. et. al, "Present: An ultra-lightweight block cipher," in *Cryptographic Hardware and Embedded Systems-CHES 2007: 9th International Workshop, Vienna, Austria, September 10-13, 2007. Proceedings 9*. Springer, 2007, pp. 450–466.

[21] T. Suzaki, K. Minematsu, S. Morioka, and E. Kobayashi, "Twine: a lightweight block cipher for multiple platforms," in *International Conference on Selected Areas in Cryptography*. Springer, 2012, pp. 339–354.

[22] K. Shibutani, T. Isobe, H. Hiwatari, A. Mitsuda, T. Akishita, and T. Shirai, "Piccolo: an ultra-lightweight blockcipher," in *Cryptographic Hardware and Embedded Systems–CHES 2011: 13th International Workshop, Nara, Japan, September 28–October 1, 2011. Proceedings 13*. Springer, 2011, pp. 342–357.

[23] J. Lu, J. Li, Z. Chen, and Y. Li, "Cryptanalysis of a type of white-box implementations of the sm4 block cipher," *The Computer Journal*, vol. 67, no. 5, pp. 1663–1673, 2024.

[24] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, and L. Wingers, "The simon and speck lightweight block ciphers," in *Proceedings of the 52nd annual design automation conference*, 2015, pp. 1–6.

[25] A. Baysal and S. Şahin, "Roadrunner: A small and fast bitslice block cipher for low cost 8-bit processors," in *Lightweight Cryptography for Security and Privacy: 4th International Workshop, LightSec 2015, Bochum, Germany, September 10–11, 2015, Revised Selected Papers 4*. Springer, 2016, pp. 58–76.

[26] R. Ankele and E. List, "Differential cryptanalysis of round-reduced sparx-64/128," in *International Conference on Applied Cryptography and Network Security*. Springer, 2018, pp. 459–475.

[27] R. Anderson, E. Biham, and L. Knudsen, "Serpent: A proposal for the advanced encryption standard," *NIST AES Proposal*, vol. 174, pp. 1–23, 1998.

[28] A. Biryukov, C. De Canniere, and G. Dellkrantz, "Cryptanalysis of safer++," in *Annual International Cryptology Conference*. Springer, 2003, pp. 195–211.

[29] J. Nechvatal, E. Barker, L. Bassham, W. Burr, M. Dworkin, J. Foti, and E. Roback, "Report on the development of the advanced encryption standard (aes)," *Journal of research of the National Institute of Standards and Technology*, vol. 106, no. 3, p. 511, 2001.

[30] C. H. Lim and T. Korkishko, "mcrypton–a lightweight block cipher for security of low-cost rfid tags and sensors," in *International workshop on information security applications*. Springer Berlin Heidelberg, 2005, pp. 243–258.

[31] J. Chen and K. Jia, "Improved related-key boomerang attacks on round-reduced threefish-512," in *International Conference on Information Security Practice and Experience*. Springer, 2010, pp. 1–18.

[32] J. Guo, T. Peyrin, A. Poschmann, and M. Robshaw, "The led block cipher," in *Cryptographic Hardware and Embedded Systems–CHES 2011: 13th International Workshop, Nara, Japan, September 28–October 1, 2011. Proceedings*. Springer Berlin Heidelberg, 2011, pp. 326–341.

[33] B. et. al, "Prince–a low-latency block cipher for pervasive computing applications," in *Advances in Cryptology–ASIACRYPT 2012: 18th International Conference on the Theory and Application of Cryptology and Information Security, Beijing, China, December 2-6, 2012. Proceedings 18*. Springer, 2012, pp. 208–225.

[34] S. Rasoolzadeh, Z. Ahmadian, M. Salmasizadeh, and M. R. Aref, "Total break of zorro using linear and differential attacks," *Cryptology ePrint Archive*, 2014.

[35] Y. Dai and S. Chen, "Cryptanalysis of full pride block cipher," *Cryptology ePrint Archive*, 2014.

[36] W. Zhang, Z. Bao, D. Lin, V. Rijmen, B. Yang, and I. Verbauwhede, "Rectangle: a bit-slice lightweight block cipher suitable for multiple platforms," *Cryptology ePrint Archive*, 2014.

[37] M. S. Turan, K. A. McKay, D. Chang, J. Kang, and J. Kelsey, "Ascon-based lightweight cryptography standards for constrained devices: Authenticated encryption, hash, and extendable output functions," NIST, Tech. Rep. NIST SP 800-232.ipd, 2023, initial Public Draft. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-232.ipd

[38] M. et. al, "Chaskey: an efficient mac algorithm for 32-bit microcontrollers," in *Selected Areas in Cryptography–SAC 2014: 21st International Conference, Montreal, QC, Canada, August 14-15, 2014, Revised Selected Papers 21*. Springer, 2014, pp. 306–323.

[39] A. R. Raza, K. Mahmood, M. F. Amjad, H. Abbas, and M. Afzal, "On the efficiency of software implementations of lightweight block ciphers from the perspective of programming languages," *Future Generation Computer Systems*, vol. 104, pp. 43–59, 2020.

[40] S. et. al, "Analysis of practical application of lightweight cryptographic algorithm ascon," in *Lightweight Cryptography Workshop 2022*. National Institute of Standards and Technology (NIST), Computer Security Resource Center (CSRC), 2022.

[41] B. Gigerl, F. Mendel, M. Schläffer, and R. Primas, "Efficient Second-Order Masked Software Implementations of Ascon in Theory and Practice," 2024, publication info: Preprint. [Online]. Available: https://eprint.iacr.org/2024/755

[42] J. Daemen, C. Dobraunig, M. Eichlseder, H. Groß, F. Mendel, and R. Primas, "Protecting against statistical ineffective fault attacks," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pp. 508–543, Jun 2020.

Raising awareness in protecting from RFID attacks

1st Andrej Vuchkovikj
School of Computer Science and Information Technology
University American College Skopje
Skopje, R.N. Macedonia
andrejvuchkovikj@outlook.com

2nd Sanja Apostolovikj
School of Computer Science and Information Technology
University American College Skopje
Skopje, R.N. Macedonia
sapostolovikjuacs@gmail.com

3rd Nikola Kostadinov
School of Computer Science and Information Technology
University American College Skopje
Skopje, R.N. Macedonia
nikolakostadinov2004@gmail.com

4th Viktor Denkovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, R.N. Macedonia
viktor.denkovski@uacs.edu.mk

5th Marija Stankova Medarovska
School of Computer Science and Information Technology
University American College Skopje
Skopje, R.N. Macedonia
marija.s.medarovska@uacs.edu.mk

6th Goce Gavrilov
School of Computer Science and Information Technology
University American College Skopje
Skopje, R.N. Macedonia
goce.gavrilov@uacs.edu.mk

***Abstract*—RFID technology is a new technology that connects the virtual and physical worlds via wireless electromagnetic fields to track and identify persons, objects, and animals. The history of RFID dates back to WWII; however, it took several decades for the technology to develop from basic operational purposes to the present day. However, the rise of RFID technology poses serious issues for cybersecurity. Spoofing involves an attacker masquerading by transmitting the signal of a tag to gain access. Unauthorized reading of specific tags happens out of contact, at times several meters away, meaning that even the most sensitive information can be leaked. In addition, some RFID tags never power off, which means they can be read at any time and for any reason, complicating the privacy problem even further. The notion that RFID advancements occur at a similar pace as other emerging technologies like IoT, AI, and even 5G presents a line of danger, deficiencies, breaches, and potential exploitation. Therefore, this research paper will address the questions posed by assessing the basic elements of RFID technology, the tangible and logistical function of RFID technology, the fields that employ RFID technology, and related cybersecurity intrusions. Ultimately, conclusions will be drawn that a more intimate relationship with RFID technology will allow for appropriate protection from RFID.**

Key*words—*RFID security, cybersecurity vulnerabilities, wireless identification technology, tag spoofing, unauthorized data access, privacy concerns, critical infrastructure protection

I. Introduction

In the rapidly changing era of digitisation, Radio Frequency Identification (RFID) was invented as a cutting-edge technology that links the digital and physical worlds. RFID can be described as a form of wireless communication that uses electromagnetic fields to detect and track everything from objects to people and animals [1]. Although RFID was originally envisioned for aircraft identification during World War II [2], the technology has advanced swiftly over the decades, and it is very helpful in today's technologically advanced society. Modern RFID systems offer a smooth, contactless data transmission method that is different from the old ways and has a much more widespread adoption. The versatility that RFID technology offers has led to transformative implementations across many sectors, from Transportation to Supply Chain Management, Healthcare and so on. From a more technical perspective, RFID systems consist of three primary components: tags, readers, and backend database systems [3]. The tags are made up of microchips and antennas that store and transmit identification data when activated by a reader's electromagnetic field. These tags can be passive, which are powered by the reader's signal; semi-passive, which are battery-assisted; or active, where the tag is self-powered, each presenting its particular security implications [4]. Readers emit radio waves and capture the data transmitted by the tags, while the backend systems process this information [5]. In the context of cybersecurity, this paper mainly focuses on increasing awareness of RFID technology because, despite its many benefits, the growth of these systems introduces significant security challenges later on. A few examples include spoofing attacks [6], which allow malicious users to replicate tags and get unauthorised access to use them, and unauthorised tag reading [7], which can happen without physical contact and expose private information to attackers from a distance of several meters. Often, because of the persistent nature of many RFID tags, they remain active and readable long after their initial purpose is fulfilled, posing as one more of the many privacy risks. As RFID technology continues to evolve through integration with innovations that are complementary to it, like the Internet of Things, Artificial Intelligence and 5G networks, the threats, security implications and attack vectors that need to be mitigated expand and evolve as well [8]. Having in mind that, security and privacy measures are crucial to be tackled, especially in areas like healthcare, where information is critical to be safeguarded [9, 10]. The objective of this paper is to emphasise and address these concerns and the increasing need to fully understand and safeguard RFID systems, given that they are vital for everyday operations. Therefore, by examining this technology and its key infrastructures and by looking at their technical operations, diverse applications, and related cybersecurity threats, this study aims to contribute to raising awareness in protecting from RFID attacks.

This paper is organized into five sections, of which the first section introduces the topic at hand, and the second section elaborates on the materials and methods used to conduct this research. The third section presents the results of this study, and in the fourth section, a discussion of the results is presented. Finally, section five concludes the article.

II. Materials And Methods

A. Used tools and technologies

For this research, Kali Linux 2024.4 was used as a specialized operating system designed for cybersecurity tasks [11]. This version with kernel 6.5.0 and Xfce 4.18.4 provided a stable version for cyber operations. It was used for various

penetration testing activities, including building and executing payloads and establishing reverse shell connections for remote access.

To allow for the adaptation of the RFID software for use, since there are many methods on how you can run a Linux virtual environment if you are not running it natively Kali Linux was the perfect operating system because of its vast security tools and features that are most appreciated in the pen-testing world, as well as ensuring the maximum compatibility with the device used in this study. To serve as a scanner and reader for this project, the PROXMARK 3 (PM3) Easy was nominated amongst many others, giving room to a flexible yet affordable device. Proxmark originated as a creation for a PhD project by Jonathan Westhues to open the door for the research of RFID systems [12]. At the time of its design, microcontrollers could not provide the high-bandwidth signal processing and precise timing required by the RFID protocols. Likewise, Software Defined Radio, or SDR, was not yet an accessible technology, requiring an entirely custom solution to be built. As such, a split architecture was designed: a microcontroller provides high-level functionality while field-programmable gate arrays (FPGA) handle the heavy lifting.

To use the PM3 easily, it was required to have it set up with all its needed firmware. The firmware that the PM3 uses is called IceMan. The well-rounded software and somewhat user-friendly version is currently managed by Chris "Iceman" Herrmann. The retrieval of this firmware can be found on GitHub - RfidResearchGroup/proxmark3: Iceman Fork - Proxmark3 [13]. To ensure a stable and seamless connection, another key aspect of the PM3 is its bootloader. The bootloader is a small chunk of code that gets the PM3 hardware up and running, and then it executes the application's "full image" firmware. For the most part, you can update the bootloader and full image all in one go if you are simply updating your firmware to the next minor revision on the same fork, however, if you are changing forks from say the factory firmware over to Iceman, then it's important to attempt to update just the bootloader first, then the application firmware. Doing so can be tricky, of course, hence, this study went through the trouble of jumping forks over to Iceman. The bootloader also determines how your PM3 communicates over USB, after which, the PM3 has gone through a few iterations when it comes to USB communications (Fig. 1).

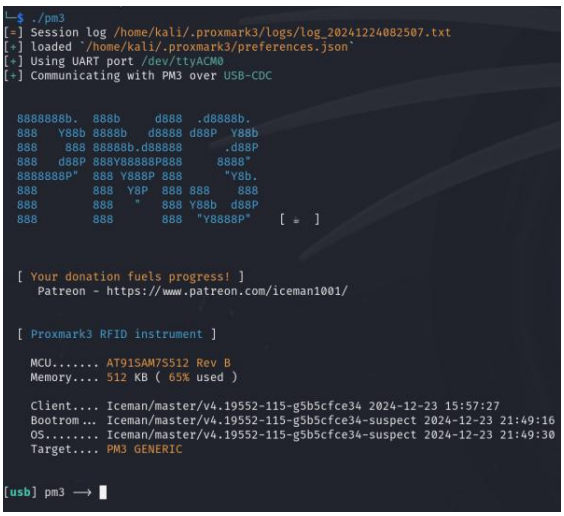


Fig. 1. Proxmark3 firmware interface

Originally, the PM3 communicated over USB using a direct protocol that required drivers to be installed for the computer to communicate with the PM3 hardware. That was eventually changed, so the USB hardware on the PM3 now acts like a virtual communication (COM) port using a very common human interface device (HID) driver interface, so no special drivers were necessary. It was required to just plug it in, and the hardware should be recognized immediately as a COM port. To ensure that a successful copy from a key card would be received, a dummy (clean card) that is compatible with the chip was necessary, and that would be scanned. There is a wide variety of RFID cards, however, there are a couple of more famous ones which usually operate on the same frequency. The only more advanced cards would be credit/debit cards and some employee ID cards. For testing and initializing this attack, a key fob for entrance doors was used, as well as a clean keycard that comes with the PM3 easily.

B. Build and Attack Approach

Using such an RFID scanner/writer can create duplicates of many RFID chips operating on the same frequency. Opening the door to a number of future vulnerabilities. This allows certain attackers to gain access to a number of restricted areas. A portable reader/writer combined with an operating system like Kali Linux, designed for expert penetration testing, makes the perfect tool to take control of everyday access to a lot of resources.

III. Results

To provide a detailed explanation of how the attack operated, a few stages had to be executed. The attack initiates with pivotal stages, meticulously arranged for seamless execution and optimal effectiveness. First, it needs to prepare the PM3 Easy. This is a crucial step in the development phase. It needs to check whether the firmware is up to date. This can be done by plugging the PM3 into one of the USB ports. After discovering which COM port it listens to, it uploads the .sh script, which is a shell script modified for writing to the PM3 easily. After the PM3 has been properly set up, the next step is to proceed with grabbing the Iceman firmware from GitHub. One step for improved work is to create a working directory so as not to confuse the cloned repository. It can be done simply by using the command (1):

$$sudo\ mkdir\ (directory_name) \quad (1)$$

After this is done, what is needed is to allow that directory to have write permissions. This is achievable by typing the following command (2):

$$sudo\ chmod\ -R\ 777\ (directory_name) \quad (2)$$

After creating the working directory, navigation is the next step by typing the command (3). Once navigated to the correct directory, the IceMan firmware can be cloned.

$$cd\ /home/(machine_name)/Desktop/(dir_name) \quad (3)$$

The following command (4) is entered in the terminal window:

$$git\ clone\ https://github.com/RfidResearchGroup/proxmar3.git \quad (4)$$

Once the directory has been cloned, it needs to be navigated into. This is done by typing the command (5). Here, the source files can be found, and the compiled versions live after the compile command is run.

cd proxmark3 (5)

After jumping into the directory before compiling, it needs to make sure the firmware is going to compile for the correct hardware. As mentioned before, there are many versions of hardware and many versions of firmware, and for the most part, they are interchangeable. The Iceman firmware was made specifically for the PM3 RDV4 hardware version, so before compiling it needs to tell it that they are not using that hardware version. To do this, it was required to create a Makefile. platform file and set a value inside. The upside of this is not needing to make a completely new file from scratch. It can just be used with the pre-existing one, make a copy of it and then edit it. To do that, use the cp command. Something worth mentioning is that Linux is case-sensitive; as such, the exact case letters for the command must be used. Knowing this, the following command is executed (6):

cp Makefile.platform.sample Makefile.platform (6)

Once the copy has been created, the file contents can be viewed with any text editor of choice. For this project, the nano text editor was used.

If Fig. 5 is examined, it can be noticed that the lines where it's described as "PLATFORM=PM3RDV4 #PLATFORM=PM3GENERIC" and the step here is simple: all that is needed to be done is to swap the hashtags of the two statements, save the file (CTRL-S) and close the nano text editor.

The compile commands are sorted together in Linux in a simple make command. All that's left to do is compile the firmware and client software by using (7).

make clean && make -j8 all (7)

This command compiles the firmware and client software and usually takes some time to complete. After the make build process is done, new files and folders will appear, a collection of PM3 flash files and a client folder as well. When PM3 is connected and showing as a virtual COM port in the device manager, it's time to flash the new firmware updates (Fig. 2).

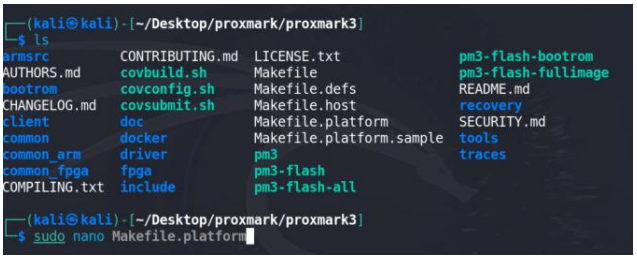


Fig. 2. The required code needs to be compiled for PM3 to function properly

Firstly, start with the bootloader by typing (8):

./PM3-flash-bootrom (8)

After that has finished, it needs to install the full image to the PM3 by entering the following command (9):

./PM3-flash-fullimage (9)

By finishing the installation, it can run the client itself by entering the command (10):

PM3 or./PM3 (10)

The final output should be something like Fig. 1.

The fun part of this project would be the scanning and cloning of the cards. It is advisable to check the functioning of the antennas. While the client is running, it can enter the command (11):

hw tune (11)

By running this command, two important lines can be seen that state "[+] LF antenna is OK" and "[HF antenna is OK]". This states that the two frequency antennas are working properly. As well as there are a few more methods on how to check if the antennas are receiving some sort of signal, such as running the command (12):

if tune (12)

After the command is run, it is required to bring some sort of RFID chip to determine whether they are reading the frequency correctly or not. If yes, then something like the following: "[=] 25927 mV / 25 V / 25 Vmax" should appear. If this is received, then this means that the PM3 Easy is functioning properly and is ready to start reading frequencies. Once it has figured out what type of frequency the chip operates on, the following command (13) can be used:

If search (13)

This command searches the contents of the chip that is being cloned. Information such as this involves finding out the chip's unique ID, which is the most important part of the data stored. When it performed the scan of the key FOB, the specific chip, which was the EM410X, was presented. This chip is of a read-only type, which operates on a low frequency of 125 kHz. They used a fixed unique ID programmed at the time of manufacture, which makes it easier to copy. After the initial read, the unique ID can be found, which is 2E006F9E16 (Fig. 3).

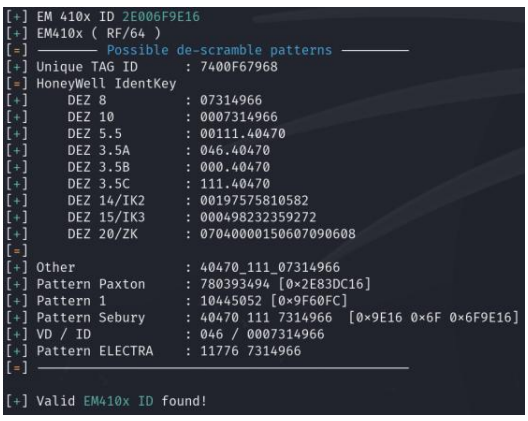


Fig. 3. Successful copying of a unique key to a dummy card

The ID 2E006F9E16 is a 40-bit identifier typically broken down into components:

- Manufacturer Code (10 bits): Identifies the manufacturer or chip series.
- Unique Serial Number (30 bits): Unique to the tag and distinguishes it from others.
- Together, they form the 40-bit ID: 2E006F9E16 (in hexadecimal format).
- In binary:
  - 2E = 00101110
  - 00 = 00000000
  - 6F = 01101111
  - 9E = 10011110
  - 16 = 00010110

Once all this information is gathered, all that's left to do is write to the new empty card. The process for this is simple. Just replace the FOB with the card on which someone wants to write the contents of the FOB. To achieve this, the final command is executed (14):

If em 410x clone -id 2E006F9E16 (14)

Once this command has been given to the firmware, and if everything works, a functioning copy of the key FOB will be enabled. To check this, the previous command (13) is used. Finally, if the unique IDs of the chips are compared and they are the same, a successful cloning of the chips is executed.

IV. Discussion

Had the RFID cloning vulnerability succeeded and all success criteria met, a malicious backdoor would be inserted into the victim's restricted areas. The hacker could go wherever they wanted since it would appear as if authorization had been granted. Nobody would be suspicious of a locked gym door or a secured terminal being accessed. With this possibility, the attacker can access various devices on the premises and steal valuable information by using additional tools for that purpose, such as USB hacking devices [14]. This was accomplished with a PM3 Easy unit, which, when coupled with open-source IceMan firmware with an array of features, is even more effective. However, there are strict success criteria. For a hacker to successfully execute this attack, he or she must be within feet of a victim's badge to clone it, operating within the operational parameters of certain RFID technology. This may be an issue because victims should always be aware of their surroundings, and even worse, it is assumed that when not in use, they use signal polarization sleeves for their badges. However, should the criminals get close, the clone itself takes mere seconds to create. The PM3 does all the necessary steps to read and clone the credential data, as evidenced in the Results section.

The vulnerability applies to all RFID devices operating at proper frequencies, but its success is relative to additional security situated surrounding the final execution. First, it was important to know the exact card type and frequency because, without these supporting criteria, the cloning would never have succeeded. Second, existing security today makes cloning vulnerabilities exceptionally hard to achieve. For instance, many more devices are now equipped with encrypted RFID chips, challenge-response authentication systems, and frequency hopping as potential scams to avoid cloning success. There are also RFID intrusion detection systems that may notify the end-user or administrator that an unwarranted read has been achieved based on the unique electromagnetic signature each time it reads.

V. Conclusion

Although RFID technology offers ease and productivity now that it is available, the safety effects and its flaws must be considered as well. The cloning of RFID badges happens at an alarming rate. These badges include specific identification cards, many hotel passkeys, or residence entry tags. Someone could just walk by the victim in a hallway, and they might accidentally give people nearby an exact copy of the key to where they are staying. This situation usually renders someone without power for some amount of time because they can generally get used to the "original" within their misuse of a copied key, which highlights a need for more strict security protocols.

These systems must run safely and securely. Encryption, strong authentication protocols, and social engineering awareness are all necessary to minimize weaknesses. Yet, global citizens are less safe, and their first line of defence is always the awareness of their surroundings. In addition to this precaution regarding international cybersecurity, those using RFID devices must be aware of when and where they are using their devices. In public areas such as shopping malls, airports, and train stations, it is more likely that someone else is attempting to scan for illicit reasons. When not in use, people can protect themselves and their RFID by powering them off or using specialized wallets that jam the signal.

As time goes on, the integration of RFID with other technologies that seem highly complementary might bring clear benefits along with inevitable drawbacks. AI use alongside solid encryption might address several weak points while mostly increasing skills. Nevertheless, throughout the many separate integrations, we introduce many potential weak points. These flaws must be carefully and thoroughly assessed as well as definitively and completely resolved.

Ultimately, in the end, faster living is undeniable with RFID technology. But if that's the case, it's a breaking point for continuing life technology if no attention to security risks is provided. Benefits must be weighed against the security risks of a technology already rooted in society, only to dig in deeper. Security concerns should not be reactive but active in the technology's planning and implementation stages from day zero. Companies need to spend time and resources on trusted card solutions with the most potential to be encrypted; average citizens need to be cautious of what cards they have and note their surroundings; governmental solutions must create standards and policies; awareness of concerns and security solutions can come from solutions by employers and schools. Technology will only grow, and whether projections in one, two, or ten years will come to a realization, security implementation should be part of the plan. However, the longer the wait to get a solid answer about RFID security concerns, the more vulnerable the system becomes. In the present, security concerns are overwhelming, but in the long run, they're manageable. As long as an understanding that RFID exists and that sometimes the convenience of its usage

should not overshadow the population's safety, everyone has everything to gain from this technology.

REFERENCES

[1] Duroc, Y., and Tedjini, S. "RFID: A Key Technology for Humanity." Comptes Rendus Physique, vol. 19, no. 1-2, Jan. 2018, pp. 64–71, https://doi.org/10.1016/j.crhy.2018.01.003.

[2] Kineber, A. F., Oke, A. E., Elseknidy, M., Hamed, M. M., and Kayode, F. S. "Barriers to the Implementation of Radio Frequency Identification (RFID) for Sustainable Building in a Developing Economy." Sustainability, vol. 15, no. 1, Jan. 2023, p. 825, https://doi.org/10.3390/su15010825.

[3] Unhelkar, B., Joshi, S., Sharma, M., Prakash, S., Mani, A. K., and Prasad, M. "Enhancing Supply Chain Performance Using RFID Technology and Decision Support Systems in the Industry 4.0–a Systematic Literature Review." International Journal of Information Management Data Insights, vol. 2, no. 2, 2022, https://doi.org/10.1016/j.jjimei.2022.100084. sciencedirect.

[4] SSolar, H., Beriain, A., Berenguer, R., Sosa, J., and Montiel-Nelson, J. A. "Semi-Passive UHF RFID Sensor Tags: A Comprehensive Review." IEEE Access, vol. 11, 2023, pp. 135583–99, https://doi.org/10.1109/access.2023.3336761. Accessed 21 Jan. 2025.

[5] Saadi, H., Touhami, R., and Yagoub, M. C. E. "Automatic Identification and Data Capture Techniques by Radio-Frequency Identification (RFID) Tags." Shaping the Future of ICT, CRC Press, Sept. 2017, pp. 63–76, https://doi.org/10.1201/9781315155241-5. Accessed 18 Mar. 2025.

[6] Wang, Y., Shen, J., Guo, X., and Dong, W. "Research on RFID Attack Methods." 2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Nov. 2020, pp. 433–37, https://doi.org/10.1109/auteee50969.2020.9315712.

[7] Halevi, T., Li, H., Ma, D., Saxena, N., Voris, J., and Xiang, T. "Context-Aware Defenses to RFID Unauthorized Reading and Relay Attacks." IEEE Transactions on Emerging Topics in Computing, vol. 1, no. 2, Dec. 2013, pp. 307–18, https://doi.org/10.1109/tetc.2013.2290537. Accessed 6 May 2021.

[8] Khan, S. I., Ray, B. R., and Karmakar, N. C. "RFID Localization in Construction with IoT and Security Integration." Automation in Construction, vol. 159, Mar. 2024, p. 105249, https://doi.org/10.1016/j.autcon.2023.105249.

[9] Denkovski, V., Stojmenovska, I., Gavrilov, G., Radevski, V. and Trajkovik, V. Investigating Privacy and Security Concerns in a Running EHealth Information System. Sept. 2023, https://doi.org/10.1109/meditcom58224.2023.10266398.

[10] Denkovski, V., Stojmenovska, I., Gavrilov, G., Radevski, V. and Trajkovik, V. "Exploring Current Challenges on Security and Privacy in an Operational EHealth Information System." Advances in Science Technology and Engineering Systems Journal, vol. 9, no. 2, Advances in Science, Technology and Engineering Systems Journal (ASTESJ), Apr. 2024, pp. 45–53, https://doi.org/10.25046/aj090206.

[11] Kali Linux 2024.4 Release (Python 3.12, Goodbye I386, Raspberry Pi Imager & Kali NetHunter) | Kali Linux Blog. Kali Linux, Kali Linux 2024.4 Release (Python 3.12, Goodbye i386, Raspberry Pi Imager & Kali NetHunter) | Kali Linux Blog, 16 Dec. 2024, www.kali.org/blog/kali-linux-2024-4-release/.

[12] Proxmark 3 | Proxmark. (n.d.). Proxmark.com. https://proxmark.com/proxmark-3-hardware/proxmark-3

[13] Iceman Fork - Proxmark3. (2023, September 5). GitHub. https://github.com/RfidResearchGroup/proxmark3

[14] Vuchkovikj, A., Kostadinov, N., Denkovski, V. and Gavrilov, G. "Accessing a Windows Machine Using a Makeshift Rubber Ducky." CIIT 2024, 19 Apr. 2024.

Moodle-GPT Chatbot: An AI Moodle Integration for Collaborative Learning

1st Nikola Kostadinov
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
nikolakostadinov2004@gmail.com

2nd Narcis Nacev
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
narcis.karanfilov@gmail.com

3rd Ekaterina Cvetkova
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
ekaterinacvetkova.biz@gmail.com

4th Andrej Vuchkovikj
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
andrejvuchkovikj@outlook.com

5th Marija Stankova Medarovska
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
marija.s.medarovska@uacs.edu.mk

6th Viktor Denkovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
viktor.denkovski@uacs.edu.mk

***Abstract*—As generative artificial intelligence (GAI) continues to advance in education, nearly achieving a true one-to-one teacher-student ratio, it also raises significant concerns about the potential extinction of traditional learning methods. Because of the non-judgmental environment AI provides, many students hesitate to ask questions in front of their peers, relying instead on AI's ability to deliver accurate answers and clear explanations. The rise of AI and its development in personal tutoring cannot and should not be stopped, but the decline of group learning, friendly debates, and collaborative discussions can be delayed. This paper introduces a method for integrating Moodle, a widely used learning management system, with generative AI tailored to a specific course's materials. This approach not only suggests but also actively promotes group discussions. The outcome of this research is a platform where students can ask questions and receive nearly instant responses, with the assurance that any inaccuracies will be corrected quickly. At the same time, this enables other students to learn from and contribute to the discussion.**

Keywords—Teaching assistant chatbot, AI, Generative AI, Moodle Forum, Collaborative Learning

I. Introduction

With the recent advances in artificial intelligence (AI), AI technology has become deeply integrated across all science and everyday life fields, fundamentally transforming how we access and process information. However, this rapid change has also led to unintended consequences, notably a decline in interpersonal communication skills among students. Recent studies suggest that extensive reliance on AI-driven educational technologies may negatively affect adolescents' social adaptability, reducing their essential interpersonal skills [1]. As we increasingly rely on AI-driven search engines regarding academic needs, traditional methods of engaging in discussions and collaborative inquiries with peers have diminished.

In response to this, this study proposes the integration of an AI-powered chatbot within the Moodle learning management system [12] as a promising solution to stimulate open dialogue and community engagement. This paper introduces Moodle-GPT, which is designed to offer a supportive educational environment where students are encouraged to share insights, ask questions, and debate course material openly. By enabling real-time conversations, the chatbot improves understanding of course material while encouraging collaborative learning, thus reducing the isolation often seen in digital education [2].

Moreover, this approach addresses the growing anxiety experienced by many students by providing a non-judgmental, accessible space for inquiries [3]. The system's public and transparent response mechanism benefits students and enables instructors to participate actively, offering expert guidance and clarification to enhance the learning experience.

Overall, the goal of this project was to bridge the gap between technological advancement and human interaction in academic settings. Moodle-GPT chatbot would serve as a stimulant for re-establishing the importance of dialogue and collaborative learning in an environment where there are capabilities of utilizing new technological advancements to personalized advantage and terms.

This study is structured in 6 sections, starting with section one that introduces the study. Section two presents related work done regarding this research. In section three, the materials and methods used for the development of the system were elaborated, and section four presents the results of the research. Section five provides a discussion of the obtained results, and finally, section six concludes the paper.

II. Related Work

Recently, AI-chatbots have been gaining significant attention regarding their usage in educational contexts due to their notable role in improving student learning efficiency with personalized academic support and engagement [4]. Integrating adaptive AI within certain learning management systems (LMS), in particular the Moodle platform, has shown promising results concerning student satisfaction and performance.

Chen and Anyanwu [5] evaluated the effectiveness of Moodle-AI chatbots in combination with metacognitive teaching approaches. Their findings emphasised the positive impact of these chatbots on students' academic performance, highlighting the effectiveness of AI-enhanced environments accompanied with conventional educational methodologies. Kaiss et al. [6] have introduced an adaptive chatbot called LearningPartnerBot integrated within Moodle, with interesting personalization techniques which adapt content according to the student's learning style, significantly improving learning outcomes. Following this,

Santana et al. [7] developed a chatbot to assist university students by answering fundamental and administrative questions. through using NLPs (natural language processing). The chatbot managed to effectively reduce administrative burdens and facilitate access to essential information. On the other hand, Sajja et al. [8] have managed to propose a curriculum-oriented intelligent educational assistant using GPT-3, designed to simplify access to course-specific content. Their platform-independent chatbot is capable of dynamically generating assistants (tailored to different academic disciplines and levels) with which the logistical demands on educators are significantly reduced. Furthermore, Liu et al. [9] have effectively presented a detailed implementation of AI tools within an introductory computer science course at Harvard University, where they were able to emphasise the capability of AI-driven teaching assistants in terms of handling diverse student queries effectively. The integration proved itself to be practical in terms of tutoring and student experiences and engagement. Additionally, Alsafari et al. [10] conducted a comparative analysis between traditional chatbots and advanced retrieval-augmented generation (RAG)-based systems in higher education. By utilising LangChain and OpenAIs GPT-3.5 Turbo model, their study demonstrated improvements and highlighted the limitations of traditional chatbots by advocating for more sophisticated AI models in educational environments.

All these studies collectively underline the significant potential of AI chatbots integrated within Moodle and similar LMS platforms. They all emphasise adaptability and personalized content delivery, along with the importance of using advanced LLMs in academic settings.

III. Materials and Methods

For the completion of this project, the following set of tools and technology were used:

- OpenAI o3-mini, which provides AI-generated responses based on course materials [11, 16];
- Moodle version 4.5.2, a free and open-source learning management system written in PHP and distributed under the GNU General Public License, was used as the base platform for the forum, allowing students to ask questions and the chatbot to answer them [12];
- Python version 3.8.10, a programming language that lets you work quickly and integrate systems more effectively, was used to create the chatbot [13, 17];
- Beautiful Soup 4.13.0, a Python library used for scraping information out of the web [14]; and
- Requests - A Python library which allows to make HTTP posts and get requests [15].

A. Theoretical Approach

Once a student posts a question on the Moodle forum, an AI retrieves the question, processes it through our designated OpenAI model via its application programming interface (API), and then posts the resulting answer back to the forum. After this, students may engage in discussions, while the professor can confirm, debunk, or enhance the provided answer.

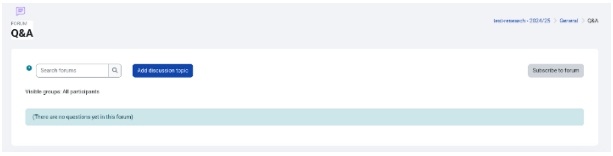


Fig. 1. The Moodle forum from a student's point of view

The idea is to have an available tutor that can answer immediately, yet still ensure the correctness of the answers given, which would allow other students to learn from the student's question with the help of a friendly discussion.

B. Practical Approach

For students to be allowed to ask questions on Moodle, the first requirement was to create a new Moodle forum. This should be done by a professor or someone who has editing privileges. The forum should be set to the "Standard forum for general use" type. If it is desired to have all students view each others' discussions, the "No Groups" Group mode should be selected. Once the Moodle forum is created, any student will be able to post a question as presented in Fig. 1.

Once the forum is created, any student can post a question by providing a subject and a message. This is a very open-ended part: the subject should be written in a generalized manner, for example, "Search Operation in an Unordered Array", while the message should contain a detailed explanation of the question and may include phrases such as "I'm confused about…" or "I can't grasp this…", although not mandatory. In most cases, a direct question would be preferable, similar to posting questions on Stack Overflow or Quora. For example, a student might ask, "Is it generally faster to find out if an item is not in the array than to find out it is?". An example of the process of posting the question is shown in Fig. 2.

After posting the question, a Python bot runs in the background, scanning for new discussions. This is achieved using the Requests library in Python to send a GET request to the Forum page. The response is in the form of an HTML code, which is further analyzed using the Beautiful Soup library in order to extract the question from the forum and forward it to the Python program responsible for analyzing and answering the question.

Once the bot detects that a discussion has been posted, generally in about 10 seconds, it retrieves the contents of the question, processes it with a model customized for the specific course material, and posts the answer as a reply to the discussion. The result is shown in Fig 3.

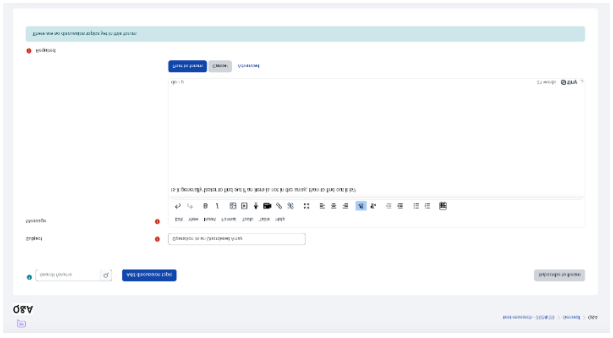


Fig. 2. Process of posting a question to the forum

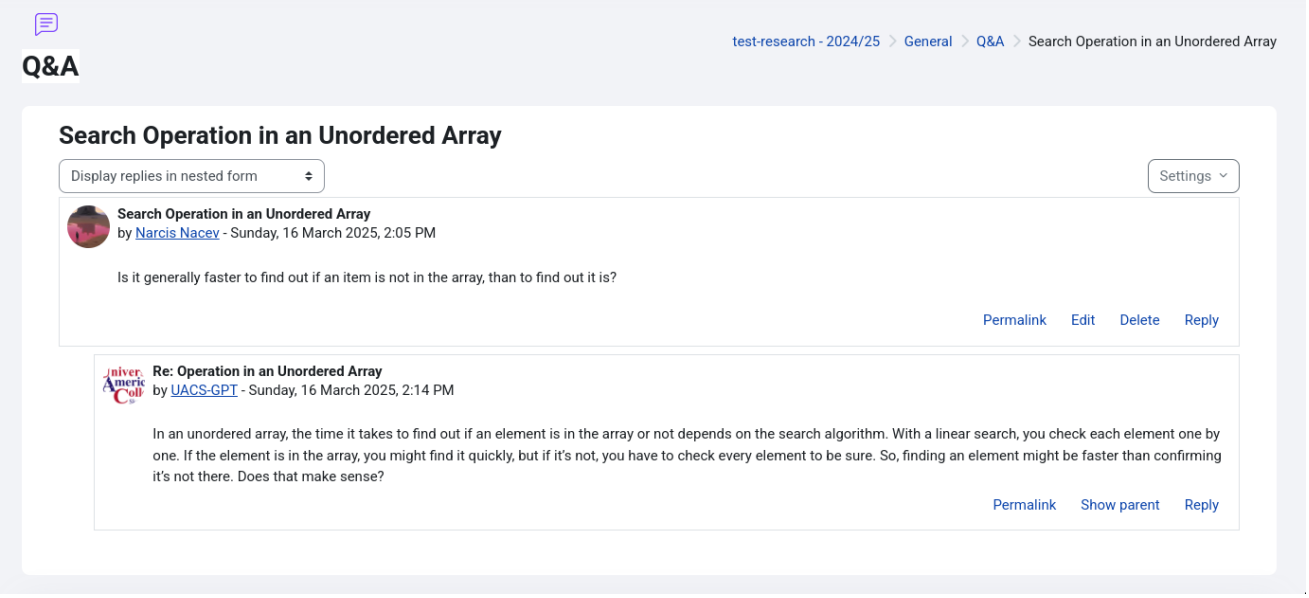


Fig. 3. Bot's reply to an asked question on the forum

C. *Analysing and Answering the Question*

This part of the program is implemented into a Python script called "UACS_GPT.py". The purpose of this file was to send an API request using the OpenAI library, with the question at hand combined with headers instructing the AI model on how to behave. For this Chatbot, the OpenAI model "gpt-4o-mini" was used.

The headers specify that the model should act as a computer science teaching assistant. It was specified to use the presentations, PDF files and further materials that were provided for each course and to answer questions from them. Furthermore, instructions like not answering the questions completely, rather giving hints towards the right answers, and being careful about academic honesty regarding complete problem solving were given. Once the program was set up and ready, the "ForumScraper.py" would call upon it, giving it the question it wanted to ask, and would send back the answer, which would get posted on Moodle.

IV. Results

The result of this research is a Moodle forum where students can initiate discussions by posting questions related to the material that is confusing them, after which the Chatbot would almost instantly reply with an AI-generated answer, allowing other students and/or professors to contribute further, enabling a collaborative discussion. The result of such a discussion is shown in Fig. 4, where the whole interaction is demonstrated. The professor's message can be emphasised with the command #UACS_GPT_STOP_ANSEWRING which is one of the predefined commands that will instruct the Chatbot to stop further replies.

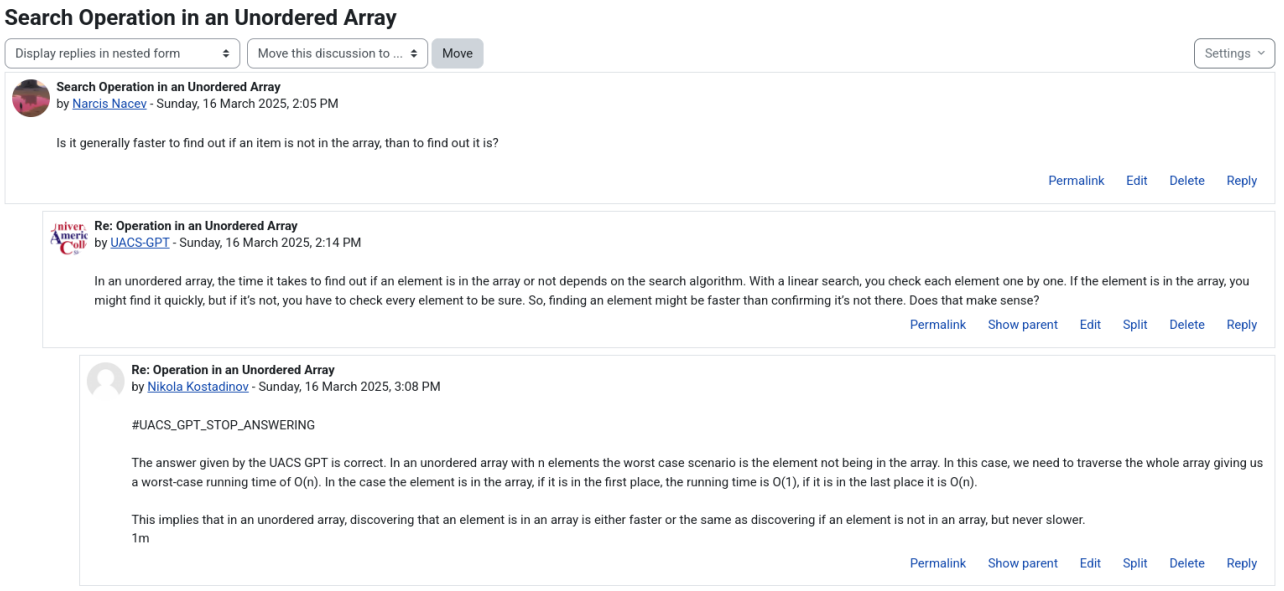


Fig. 4. Sample discussion on the forum

TABLE I. OTHER USEFUL COMMANDS FOR THE UACS-GPT MOODLE CHATBOT

Commands	Usage
#UACS_GPT_ STOP_ANSWERING	The UACS GPT will be deactivated for this discussion.
#UACS_GPT_ RESUME_ANSWERING	The UACS GPT will be activated again.
#UACS_GPT_ ANSWER_{n}	Replies to the n^{th} message from the start of the discussion. It does not count its messages.
#UACS_GPT_ TRY_AGAIN	Edit the previous answer with a newly generated answer.
#UACS_GPT_ ELABORATE	Posts a new answer with a more detailed explanation.

Furthermore, the discussion can continue back and forth with the student, stopping once the command #UACS_GPT_STOP_ANSWERING is written. This would be a case where the professor can either confirm, debunk, or further explain the answer as shown in Fig 4.

V. DISCUSSION

By incorporating a few of the AI technologies to the Moodle-GPT chatbot in Moodle, our project addresses head-on one of the major issues created by AI: diminishing students' interpersonal and collaborative communication skills [1]. This approach aligns well with recent developments highlighted in existing research, as seen in the related works section:

- Personalized Learning Experience: Similar to the LearningPartnerBot presented by Kaiss et al. [6], Moodle-GPT provides personalized assistance tailored specifically to the course materials, enhancing the educational experience through context-related interactions. However, unlike LearningPartnerBot, our chatbot distinctly prioritizes collaborative and community-based dialogues rather than focusing only on individual learning.
- Impact on Academic Performance: As demonstrated by Chen and Anyanwu [5], Moodle-integrated AI chatbots can significantly improve academic performance when paired with effective teaching methodologies. Our chatbot has the potential to boost student performance by actively engaging students in public discussions and debates, though we emphasize the additional benefit of reducing student anxiety through transparent communication.
- Teaching Assistant Capabilities: The LLM-powered teaching assistant approach proposed by Santana et al [7], demonstrates significant advancements over traditional chatbots, which aligns closely with the methodology utilized in our Moodle-GPT project. Our chatbot similarly emphasizes adaptability, effectively addressing a diverse range of student queries. Moreover, our solution uniquely emphasizes public interactions in order to encourage stronger community engagement.

Additionally, when compared with the existing Moodle plugin Corolair [8], which also utilizes AI to address student inquiries, our Moodle-GPT chatbot provides a notable advantage by explicitly promoting public discussions rather than individualized responses alone. This emphasis on open, anxiety-free discussion uniquely positions our chatbot as a solution that not only meets educational needs but also actively promotes social interaction and community-building among students and instructors.

However, our Moodle-GPT system does have certain limitations that should be acknowledged and addressed in future development, especially when it comes to continually refining the precision of generated answers to maintain high-quality educational exchanges.

In summary, the practical usage of the Moodle-GPT chatbot represents a substantial step forward in addressing educational and social isolation caused by new AI technology development. By facilitating open dialogue, reducing student anxiety, and promoting instructor participation, our chatbot effectively bridges the gap between advanced technological tools and human educational practices, which was our goal to begin with. This approach promises enhanced collaborative learning experiences, increased student engagement, and improved academic outcomes.

VI. CONCLUSION

To conclude, while striving to achieve a one-to-one teacher-student learning model, implementing a Chatbot for a Moodle forum promotes a more effective peer-to-peer learning environment. AI tools based on Large Language Models (LLMs), such as ChatGPT, Gemini, and Deepseek, can assist with clarification and answering direct questions. However, their ability to help is limited to the specific queries asked, and their responses are based on pre-existing training data, not the specific course content. An open forum, a more open environment, and supporting communication between peers can broaden students' understanding, enabling them to learn not just with the help of AI but also through insights from colleagues.

The current Chatbot implementation may offer multiple opportunities for improvement. Future research may involve changing the chatbot's infrastructure from a web scraping approach to a more advanced bot using Moodle API with event notifications. This can improve the program's speed, and security, alongside removing the unnecessary operations it runs in the background. Furthermore, the chatbot can be adapted to not only use specific materials provided by the professors but fetch all

materials from the Moodle course in real time, allowing it to help with confusion of course results, grades or wordings, making it even more of an assistant.

REFERENCES

[1] T. Lai, C. Xie, M. Ruan, Z. Wang, H. Lu, and S. Fu, "Influence of artificial intelligence in education on adolescents' social adaptability: The mediatory role of social support," PLoS ONE, vol. 18, no. 3, p. e0283170, Mar. 2023, doi: 10.1371/journal.pone.0283170.

[2] Moraes Neto, A. J., and Fernandes, M. A. "Chatbot and Conversational Analysis to Promote Collaborative Learning in Distance Education." IEEE Xplore, vol. 2161-377X, 1 July 2019, pp. 324–26, https://doi.org/10.1109/ICALT.2019.00102. Accessed 7 Jan. 2023.

[3] Oghenekaro, L. U., and Okoro, C. O. "Artificial Intelligence-Based Chatbot for Student Mental Health Support." *Open Access Library Journal*, vol. 11, no. 5, May 2024, pp. 1–14, https://doi.org/10.4236/oalib.1111511.

[4] Labadze, L., Grigolia, M., and Machaidze, L. "Role of AI Chatbots in Education: Systematic Literature Review." *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, Springer Nature, Oct. 2023, https://doi.org/10.1186/s41239-023-00426-1.

[5] Chen, H., and Anyanwu, C. C. "AI in Education: Evaluating the Impact of Moodle AI-Powered Chatbots and Metacognitive Teaching Approaches on Academic Performance of Higher Institution Business Education Students." *Education and Information Technologies*, Jan. 2025, https://doi.org/10.1007/s10639-024-13235-4.

[6] Kaiss, W., Mansouri, K., and Poirier, F. "Effectiveness of an Adaptive Learning Chatbot on Students' Learning Outcomes Based on Learning Styles." *International Journal of Emerging Technologies in Learning (Ijet)*, vol. 18, no. 13, Kassel University Press, July 2023, pp. 250–61, https://doi.org/10.3991/ijet.v18i13.39329.

[7] Santana, R., Ferreira, S., Rolim, V. B., Miranda, P. B., Nascimento, A. C., and Mello, R. F. "A Chatbot to Support Basic Students Questions." *Latin American Conference on Learning Analytics* 2021.

[8] Sajja, R., Yusuf Sermet, Cwiertny, D. M., and Demir, I. "Platform-Independent and Curriculum-Oriented Intelligent Assistant for Higher Education." *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, Open University of Catalonia, July 2023, https://doi.org/10.1186/s41239-023-00412-7.

[9] Portland, O., and Malan. *Teaching CS50 with AI: Leveraging Generative Artificial Intelligence in Computer Science Education*. ACM, 2024, https://doi.org/10.1145/3626252.3630938.

[10] Alsafari, B., Atwell, E., Walker, A., and Callaghan, M. "Towards Effective Teaching Assistants: From Intent-Based Chatbots to LLM-Powered Teaching Assistants." *Natural Language Processing Journal*, Elsevier BV, Aug. 2024, pp. 100101–1, https://doi.org/10.1016/j.nlp.2024.100101.

[11] "OpenAI O3-Mini." *Openai.com*, 2025, openai.com/index/openai-o3-mini/.

[12] "Moodle 4.5.2 | Moodle Developer Resources." *Moodledev.io*, 20 Feb. 2025, moodledev.io/general/releases/4.5/4.5.2. Accessed 31 Mar. 2025.

[13] "Python Release Python 3.8.10." *Python.org*, www.python.org/downloads/release/python-3810/.

[14] Richardson, L. "Beautiful Soup Documentation — Beautiful Soup 4.4.0 Documentation." *Crummy.com*, 2019, www.crummy.com/software/BeautifulSoup/bs4/doc/.

[15] Reitz, K. "Requests: Python HTTP for Humans." *PyPI*, 22 May 2023, pypi.org/project/requests/.

[16] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R., 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[17] Hunt, J., 2019. *Advanced guide to Python 3 programming* (pp. 35-42). Berlin: Springer.

Exploring Transformer Models for EMG Signal Classification: A Case Study in ALS Detection

Mihail Gjorgjiev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
mihail.gjorgjiev@students.finki.ukim.mk

Kostadin Mishev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
kostadin.mishev@finki.ukim.mk

Monika Simjanoska Misheva
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
monika.simjanoska@finki.ukim.mk

***Abstract*—Amyotrophic Lateral Sclerosis is a progressive neurodegenerative disease that affects motor neurons, leading to muscle weakness and eventual paralysis. Early and accurate diagnosis is crucial for timely intervention and improved patient outcomes. This study explores the potential of Machine learning and Large Language Models for ALS prediction using electromyography signals. Specifically, we evaluate XGBoost, LLaMA, DeepSeek and Gemma models in four classification experiments: (1) healthy individuals vs. ALS patients, (2) healthy individuals vs. myopathy patients, (3) ALS patients vs. myopathy patients, and (4) healthy vs. ALS vs. myopathy patients.**

While XGBoost demonstrated the highest performance, the transformer-based models demonstrated the feasibility for neuromuscular disorder detection. Future work will focus on refining model performance and exploring hybrid approaches to enhance clinical applicability. [1]

***Index Terms*—Amyotrophic Lateral Sclerosis, ALS, Electromyography, EMG, Machine Learning, Transformers, Large Language Models, LLMs, XGBoost, LLaMa, DeepSeek, Gemma**

I. Introduction

The human neuromuscular system depends on the connection between motor neurons and muscles to enable movement. Motor neurons in the brain and spinal cord transmit electrical impulses to muscle fibers, triggering contractions that produce voluntary motion. These impulses travel through motor unit pathways, with each motor neuron controlling multiple muscle fibers. Proper function of this system is essential for coordinated movement, and any disruption can cause severe motor impairments [1].

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease that mainly affects the motor neurons, leading to muscle weakness, atrophy, and eventually, paralysis [2]. As motor neurons deteriorate, their ability to transmit electrical impulses decreases, resulting in abnormal muscle activity and loss of voluntary control. This breakdown in neuromuscular communication sets ALS apart from other neuromuscular disorders, such as myopathies, which mainly affect the muscle fibers rather than neurons themselves [3]. However, diagnosing ALS is difficult due to symptom overlap with other conditions and reliance on subjective clinical assessments.

Electromyography (EMG) plays a crucial role in diagnosing neuromuscular disorders, including ALS [4]. EMG measures the electrical activity of muscles in response to neural stimulation, providing insights into motor unit function. By analyzing EMG signals, clinicians can identify abnormalities such as spontaneous muscle activity (fasciculations), altered firing patterns, and reduced recruitment of motor units. These characteristics help differentiate ALS from other conditions like myopathies, where muscle fiber damage leads to different EMG signal patterns. Despite its diagnostic value, interpreting EMG signals requires expertise and can be time-consuming, necessitating automated approaches for improved efficiency and accuracy [5].

Advances in Artificial Intelligence (AI) and Machine Learning (ML) have enabled powerful tools for analyzing biomedical signals, including EMG. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can automatically extract features from raw EMG and classify neuromuscular disorders with high accuracy. More recently,

Funded 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.

[1]The code used in this study is available on GitHub: https://github.com/ChatMED/ALS-EMG.

transformer-based models, including Large Language Models (LLMs) like LLaMA, DeepSeek, and Gemma, have shown promise in processing sequential data. Originally developed for natural language processing (NLP), these models can capture complex temporal patterns in EMG signals, making them well-suited for detecting ALS and other neuromuscular conditions.

In this study, we explore the feasibility of using LLMs to analyze EMG signals for neuromuscular disorder classification. Specifically, we investigate their ability to differentiate in binary experimental setup (three cases), and multiclass experimental setup (among the three classes).

The rest of the paper is organized as follows: Section II reviews related work, providing context for EMG-based classification and the use of traditional ML models in neuromuscular disorder detection. Section III details the methodology, including dataset characteristics, feature extraction, and model implementation. Section III-D presents the experimental results, while Section V discusses the results implications, comparing model performance and highlighting key findings. Finally, Section VI concludes the study and outlines potential directions for future research.

II. Related Work

Recent research on biomedical signal classification indicates a progressive shift from traditional ML methods to deep learning (DL) and LLMs. Since ALS is a neurodegenerative disease, most research on its diagnosis and progression monitoring has focused on electroencephalography (EEG), with relatively limited studies exploring EMG-based classification.

Classical techniques such as Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), and K-Nearest Neighbors (KNN) have been widely applied to electromyography (EMG)-based classification tasks for neuromuscular disorders, particularly amyotrophic lateral sclerosis (ALS). These methods often rely heavily on manual feature engineering, requiring domain expertise and limiting scalability. Mokdad et al. [6] achieved 95.8% accuracy using SVM for ALS diagnosis, while Mishra et al. [7] employed an Extreme Learning Machine (ELM), reporting 88% accuracy, underscoring the importance of feature selection. Similarly, Ahmed and Islam [8] demonstrated that while artificial neural networks (ANNs) could be effective for neuromuscular disorder classification, their SVM model underperformed at 70% accuracy, again emphasizing the crucial role of optimized feature extraction. Recent advancements have steered research towards DL approaches, particularly convolutional neural networks (CNNs), which can learn hierarchical features directly from raw EMG signals. Hassan et al. [9] introduced ALSNet, a 1D CNN trained on raw EMG data, achieving 97.74% accuracy and eliminating the need for handcrafted features.

With the rise of transformer-based architectures, their application has rapidly expanded beyond NLP, with a growing body of research exploring their effectiveness across various physiological signals and medical conditions. Despite this momentum, the use of LLMs for EMG signal analysis remains largely underexplored. Existing studies have focused on distinct contexts, such as synthetic data generation [10] and motion intention recognition [11], but to date, no work has addressed the specific classification challenge investigated in this study.

III. The Methodology

Figure 1 illustrates the methodological workflow, encompassing data preprocessing, feature extraction, and data sampling stages. The processed data is then utilized in two parallel pipelines: one leveraging XGBoost and the other transformer-based LLMs, including LLaMA, DeepSeek, and Gemma. Each model undergoes training or fine-tuning, followed by a unified evaluation phase to assess their discriminative performance in classifying EMG signals.

A. Dataset

This study uses the EMGLAB2001 dataset [12], which contains EMG recordings from healthy individuals, ALS patients, and myopathy patients. The raw EMG signals were recorded from the biceps brachii muscle under various conditions. The dataset includes recordings from 22 participants, comprising 12 males and 10 females, with ages ranging from 19 to 67 years, divided into three groups: 7 ALS patients, 9 control (healthy) subjects, and 6 myopathy patients. Each recording captures the biceps brachii's electrical activity, offering valuable data for analyzing neuromuscular disorders. To ensure consistency and usability, recordings were preprocessed to remove artifacts and normalize amplitudes, improving signal quality and enhancing feature extraction and classification.

B. Feature Extraction

The EMG signals in the dataset have a sampling rate of 23,437.5 Hz and a total duration of 11.2 seconds per recording. To enhance feature extraction and analysis, signals are segmented into 250 ms intervals, ensuring that each segment contains sufficient temporal information, in terms of how muscle activity changes within a specific time window [9].

These segments are then used to compute various time-domain [13], frequency-domain [8], nonlinear, and spectral features [14], which are essential for differentiating between healthy individuals, ALS patients, and those with myopathies. To use these features for classification,

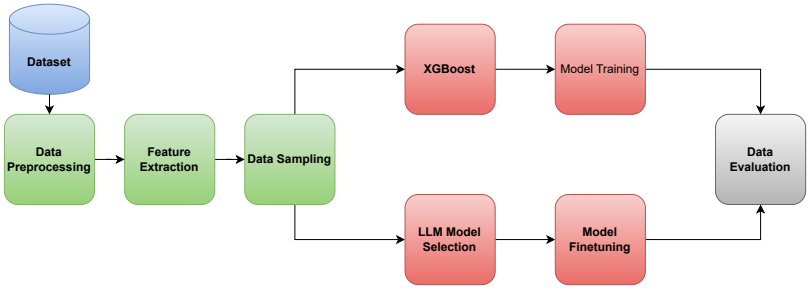


Fig. 1. The Methodology Workflow.

it is important to define their formulas and understand how each one helps distinguish neuromuscular conditions [15]:

1) Mean Absolute Value: provides an estimate of the signal's average amplitude and represents muscle activation levels. ALS patients typically show altered muscle activation due to motor neuron degeneration.

$$MAV = \frac{1}{N}\sum_{i=1}^{N}|x_i| \tag{1}$$

2) Waveform Length: measures the cumulative length of the EMG waveform, indicating signal complexity and intensity. ALS patients may exhibit increased waveform irregularities due to involuntary muscle activity.

$$WL = \sum_{i=1}^{N-1}|x_{i+1} - x_i| \tag{2}$$

3) Zero Crossing: Zero crossings count the number of times the signal changes sign. This feature helps identify abnormal firing patterns in ALS, where muscle activity may become erratic.

$$ZC = \sum_{i=1}^{N-1} f(x_i, x_{i+1}) \tag{3}$$

where

$$f(x_i, x_{i+1}) = \begin{cases} 1, & \text{if } (x_i \cdot x_{i+1} < 0) \text{ and } |x_i - x_{i+1}| > \Delta \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where Δ is a small threshold to prevent noise-induced crossings.

4) Slope Sign Change: SSC measures the frequency of slope direction changes, aiding the detection of signal complexity and instability, which is often higher in ALS patients.

$$SSC = \sum_{i=2}^{N-1} g(x_{i-1}, x_i, x_{i+1}) \tag{5}$$

where

$$g(x_{i-1}, x_i, x_{i+1}) = \begin{cases} 1, & (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) > \Delta \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where Δ is a small threshold to prevent noise-induced crossings.

5) Mean Frequency: MNF represents the average frequency of the power spectrum, helping differentiate between normal and ALS-affected muscle activity. ALS patients often show a shift in frequency due to motor unit loss.

$$MNF = \frac{\sum_{i=1}^{N} f_i P(f_i)}{\sum_{i=1}^{N} P(f_i)} \tag{7}$$

where f_i is the frequency and P(f_i) is the power at that frequency.

6) Median Frequency: MDF is the frequency that divides the power spectrum into two equal halves, reflecting neuromuscular fatigue and motor unit recruitment abnormalities in ALS.

$$\sum_{f=0}^{MDF} P(f) = \frac{1}{2}\sum_{f=0}^{f_{max}} P(f) \tag{8}$$

where P(f) is the power spectral density.

7) Approximate Entropy: quantifies the regularity of a signal. Lower values indicate more predictable patterns, while higher values reflect irregularity, as seen in ALS patients' EMG signals.

$$ApEn(m, r, N) = \phi(m, r) - \phi(m+1, r) \tag{9}$$

where m is the embedding dimension, r is the tolerance threshold, and Φ represents the average similarity of sequences in the signal.

8) Sample Entropy: measures signal complexity by evaluating pattern predictability. ALS patients tend to have reduced entropy due to impaired neuromuscular control.

$$SampEn(m, r, N) = -\ln\left(\frac{A}{B}\right) \tag{10}$$

where A and B are counts of similar sequences of length *m+1* and *m*, respectively.

9) Fractal Dimension: quantifies signal complexity and self-similarity, which can help distinguish ALS patients who exhibit irregular neuromuscular firing.

$$FD = \lim_{k \to \infty} \frac{\log L(k)}{\log(1/k)} \tag{11}$$

where *L(k)* represents the length of the curve at a given scale *k*.

10) Wavelet Entropy: quantifies the disorder in a signal's wavelet coefficients. ALS patients often show increased entropy due to abnormal neuromuscular activity.

$$WE = -\sum_{i=1}^{N} p_i \log p_i \tag{12}$$

where p_i is the normalized energy distribution.

11) Mean Instantaneous Frequency: MIF measures the average frequency content of the signal in the time-frequency domain, helping to identify neuromuscular instability in ALS.

$$MIF = \frac{\sum_{i=1}^{N} t_i f_i}{\sum_{i=1}^{N} t_i} \tag{13}$$

where f_i is the instantaneous frequency at time t_i.

C. The Models

For the purpose of this study, four models were used to classify EMG signals based on the extracted features: XGBoost, LLaMA, DeepSeek and Gemma. Each model processes the data differently, allowing us to compare traditional ML with LLM techniques.

1) XGBoost: is a decision tree-based model optimized for structured data. It combines multiple weak models to enhance predictions, making it a strong baseline for evaluating transformer-based methods.

2) LLaMA: is a transformer model originally developed for NLP but capable of handling sequential data. Instead of raw EMG signals, it processes extracted features to identify patterns associated with neuromuscular disorders [16].

3) DeepSeek: is another transformer model, is optimized for long-range dependencies and complex pattern recognition. This enhances its ability to capture subtle EMG variations relevant for ALS detection [17].

4) Gemma: is a lightweight transformer designed for efficiency. Its optimized architecture may offer better adaptability to structured data, making it a promising alternative for EMG-based classification, particularly in resource-limited settings.

D. Experiments

To evaluate model performance in distinguishing neuromuscular conditions, four classification experiments were conducted. These experiments assessed the models' ability to differentiate among control subjects, ALS patients, and myopathy patients in both binary and multiclass settings.

1) Control vs. ALS: This experiment distinguishes ALS patients from healthy individuals, determining whether the extracted features effectively capture ALS-related changes in EMG signals.

2) Control vs. Myopathy: This experiment separates myopathy patients from healthy individuals, verifying whether the extracted features reliably capture EMG signal changes caused by muscle fiber degeneration.

3) ALS vs. Myopathy: This experiment compares ALS and myopathy patients, conditions with similar symptoms but distinct pathological origins (neurogenic vs. myogenic). Accurate classification here demonstrates the models' ability to capture disease-specific EMG features.

4) Control vs. ALS vs. Myopathy: This experiment classifies all three groups simultaneously, requiring the models to distinguish normal activity, neurogenic degeneration (ALS), and myogenic degradation (myopathy). Success here indicates model robustness for real-world diagnostics.

We trained on 15 patients (5 per class) and tested on 7 (4 control, 2 ALS, 1 myopathy), ensuring class balance. Sample sizes were: Control (3300/1100), ALS (3062/1173), and Myopathy (3562/1095) for training/testing.

E. Evaluation Metrics

To compare the models' performance, four standard metrics have been used: precision, recall, F1-score and accuracy. These metrics provide a comprehensive assessment, with precision measuring the reliability of positive predictions and recall measuring the model's ability to identify positive cases. F1-score balances precision and recall, and accuracy offers an overall performance measure for total correct classified samples. [18].

IV. Results

As shown in Table I, the results indicate that XGBoost consistently achieved the highest F1-scores across all experiments, outperforming the smaller LLaMA-3.2-1B and DeepSeek models. This suggests that traditional tree-based models are more effective for classifying EMG signal features than transformer-based models, which are primarily designed for text processing. However, larger transformer models, such as LLaMA-3.1-8B and Gemma-7B, showed notable performance improvements, reaching accuracy levels comparable to XGBoost. This suggests promising potential for replacing traditional models like XGBoost in future multimodal diagnostic frameworks.

TABLE I
PERFORMANCE METRICS OF THE MODELS ACROSS THE PREDEFINED EXPERIMENTS.

Experiment	Model	Precision	Recall	F1	Accuracy
Control-ALS	XGBoost	72.56%	81.15%	76.61%	74.43%
Control-ALS	LLaMA-3.2-1B	78.10%	42.88%	55.36%	64.32%
Control-ALS	LLaMA-3.1-8B	76.60%	76.14%	76.16%	76.33%
Control-ALS	DeepSeek	75.07%	42.88%	54.58%	63.17%
Control-ALS	Gemma-7B	72.69%	73.28%	72.94%	73.12%
Control-Myopathy	XGBoost	76.15%	57.16%	65.31%	69.70%
Control-Myopathy	LLaMA-3.2-1B	65.84%	33.97%	44.81%	58.26%
Control-Myopathy	LLaMA-3.1-8B	65.78%	64.89%	64.41%	64.92%
Control-Myopathy	DeepSeek	53.49%	50.95%	52.19%	53.43%
Control-Myopathy	Gemma-7B	66.92%	66.49%	66.29%	66.51%
ALS-Myopathy	XGBoost	81.13%	92.75%	86.55%	85.09%
ALS-Myopathy	LLaMA-3.2-1B	76.23%	81.50%	78.78%	77.29%
ALS-Myopathy	LLaMA-3.1-8B	85.03%	85.05%	85.04%	85.05%
ALS-Myopathy	DeepSeek	75.12%	78.26%	76.65%	75.35%
ALS-Myopathy	Gemma-7B	84.55%	83.50%	83.59%	83.99%
Control-ALS-Myopathy	XGBoost	67.29%	65.71%	65.74%	66.03%
Control-ALS-Myopathy	LLaMA-3.2-1B	50.21%	50.98%	50.49%	52.43%
Control-ALS-Myopathy	LLaMA-3.1-8B	61.19%	60.72%	60.55%	61.13%
Control-ALS-Myopathy	DeepSeek	48.80%	50.16%	48.71%	50.47%
Control-ALS-Myopathy	Gemma-7B	60.49%	59.92%	59.70%	60.36%

For the binary classification tasks, XGBoost performed best in distinguishing between ALS and Myopathy (F1-score: 86.55%), followed by Control vs. ALS (F1-score: 76.61%) and Control vs. Myopathy (F1-score: 65.31%). The larger transformer models followed closely, with LLaMA-3.1-8B achieving F1-scores 85.04%, 76.16%, and 64.41% for ALS-Myopathy, Control-ALS, and Control-Myopathy, respectively, while Gemma-7B obtained F1-scores 83.59%, 72.94%, and 66.29% in the same tasks. The smaller models, LLaMA-3.2-1B and DeepSeek, struggled, with their lowest performance observed in Control vs. Myopathy (F1-scores: 44.81% and 52.19%).

In the multiclass classification task (Control-ALS-Myopathy), performance dropped for all models. XGBoost remained the strongest model with an F1-score of 65.74%, but LLaMA-3.1-8B and Gemma-7B followed closely at 60.55% and 59.70%. In contrast, LLaMA-3.2-1B and DeepSeek achieved 50.49% and 48.71%, highlighting the difficulty in distinguishing among the three conditions.

V. DISCUSSION

summarizes EMG classification models from prior studies using the EMGLAB2001 dataset, ranging from traditional ML methods like SVM and ELM to DL approaches such as CNNs. Performance variability reflects differences in model complexity, class configuration (binary vs. multiclass), and train-test split strategies. The best performing model among the prior works is ALSNet [9], a 1-D dilated CNN that achieved an impressive 97.74% accuracy, with F1-score of 97.67%, significantly outperforming classical methods like SVM [6] and ELM [7]. This is attributable to ALSNet's ability to process raw EMG signals without handcrafted features, exploiting local dependencies through dilation and convolution. In contrast, our experimental results (Table I) using a suite of modern transformer-based models (LLaMA, DeepSeek, Gemma) and XGBoost on the same EMGLAB2001 dataset demonstrate competitive yet modest performance, particularly in the binary classification tasks (e.g., Control vs. ALS and ALS vs. Myopathy). The LLaMA-3.1-8B model, for instance, achieved 76.33% accuracy and 76.16% F1-score on the Control-ALS task, aligning closely with XGBoost and Gemma. Interestingly, in the ALS vs. Myopathy classification, LLaMA-3.1-8B reached an accuracy of 85.05%, approaching the performance of ELM (88%) [7] and only trailing ALSNet slightly. This reinforces the hypothesis that transformer-based models are particularly effective in discerning between distinct pathological classes when enough signal differentiation exists. However, in multiclass settings (Control vs. ALS vs. Myopathy), performance of all tested models degraded significantly (e.g., max 66.03% accuracy for XGBoost). This drop mirrors trends observed in prior literature where multiclass EMG classification remains challenging due to overlapping features between disease states [7]. Moreover, the studies such as [8] and [6] used random splits for training and testing, which may lead to data leakage or inflated performance if samples from the same subjects appear in both sets. In contrast, our approach enforces consistent, predefined experiments across class splits, offering more conservative but reliable evaluation.

Overall, the performance gap between ALSNet and our transformer-based models can be explained by several

TABLE II
PERFORMANCE METRICS FOR OTHER EMG CLASSIFICATION MODELS.

Ref.	Dataset	Model	Class Split	Split	Accuracy	Precision	Recall	F1-Score
[8]	EMGLAB2001	SVM	Binary (Control vs. ALS+Myopathy)	Random	70%	-	-	-
[6]	EMGLAB2001	SVM	Binary (Control vs. ALS)	Random	95.8%	92.59%	92%	92.3%
[7]	EMGLAB2001	ELM	Multiclass (Control vs. ALS vs. Myopathy)	Random	88%	70%	96%	80.96 %
[9]	EMGLAB2001	CNN (ALSNet)	Binary (Control vs. ALS)	80/20	97.74%	96.77%	98.59%	97.67%

factors:

- Input modality: ALSNet directly consumes raw EMG waveforms, whereas our models are likely operating on transformed or feature-extracted representations.
- Architecture design: ALSNet's convolutional architecture is inherently suited for time-series biomedical signals, while transformer models, though powerful, might require architectural adaptations for EMG tasks.
- Feature extraction: ALSNet eliminates hand-crafted features, while classical models and our experimental pipeline may depend on preprocessing that limits signal richness.

These findings suggest that further tuning, especially domain specific finetuning or combining CNN and transformer architectures, could help close the performance gap. Additionally, using raw waveform inputs with large models may better leverage their ability to capture complex temporal patterns.

VI. CONCLUSION

This study evaluated XGBoost, LLaMA-3.2-1B, DeepSeek, LLaMA-3.1-8B, and Gemma-7B for classifying neuromuscular disorders using EMG-derived features. XGBoost consistently outperformed smaller transformer models, showing that traditional ML remains highly effective for structured EMG data. However, larger LLMs like LLaMA-3.1-8B and Gemma-7B achieved comparable results, indicating that increased model size boosts performance. These results underscore the importance of aligning model choice with data characteristics.

Future work could explore integrating domain-specific adaptations, or combining traditional ML with DL approaches for enhanced classification performance.

REFERENCES

[1] J. Bashford, K. Mills, and C. Shaw, "The evolving role of surface electromyography in amyotrophic lateral sclerosis: A systematic review," *Clinical Neurophysiology*, no. 4, pp. 942–950, 2020.

[2] M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, and M. C. Zoing, "Amyotrophic lateral sclerosis," *The lancet*, vol. 377, no. 9769, pp. 942–955, 2011.

[3] M. B. Bromberg, D. Forshew, K. Nau, J. Bromberg, Z. Simmons, and T. Fries, "Motor unit number estimation, isometric strength, and electromyographic measures in amyotrophic lateral sclerosis," *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, vol. 16, no. 11, pp. 1213–1219, 1993.

[4] S. Nagineni, S. Taran, and V. Bajaj, "Features based on variational mode decomposition for identification of neuromuscular disorder using emg signals," *Health information science and systems*, vol. 6, pp. 1–10, 2018.

[5] Y. Wu, M. Á. M. Martínez, and P. O. Balaguer, *Overview of the Application of EMG Recording in the Diagnosis and Approach of Neurological Disorders*, vol. 10. IntechOpen London, UK, 2013.

[6] A. Mokdad, S. M. E. A. Debbal, and F. Meziani, "Diagnosis of amyotrophic lateral sclerosis (als) disorders based on electromyogram (emg) signal analysis and feature selection," *Polish Journal of Medical Physics and Engineering*, vol. 26, no. 3, pp. 155–160, 2020.

[7] V. K. Mishra, V. Bajaj, and A. Kumar, "Classification of normal, als, and myopathy emg signals using elm classifier," in *2016 2nd international conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB)*, pp. 455–459, IEEE, 2016.

[8] T. Ahmed and M. K. Islam, "Emg signal classification for detecting neuromuscular disorders," in *Journal of Physics: Conference Series*, vol. 1921, p. 012043, IOP Publishing, 2021.

[9] K. N. Hassan, M. S. A. Hridoy, N. Tasnim, *et al.*, "Alsnet: a dilated 1-d cnn for identifying als from raw emg signal. icassp 2022–2022 ieee international conference on acoustics, speech and signal processing (icassp)," 2022.

[10] J. J. Bird, M. Pritchard, A. Fratini, A. Ekárt, and D. R. Faria, "Synthetic biological signals machine-generated by gpt-2 improve the classification of eeg and emg through data augmentation," *IEEE Robotics and Automation Letters*, pp. 3498–3504, 2021.

[11] L.-L. Li, G.-Z. Cao, H.-J. Liang, Y.-P. Zhang, and F. Cui, "Human lower limb motion intention recognition for exoskeletons: A review," *IEEE Sensors Journal*, pp. 30007–30036, 2023.

[12] M. Nikolic, C. Krarup, and J. A. Sørensen, "Detailed analysis of clinical electromyography signals," *University of Copenhagen, the Faculty of Health Science*, 2001.

[13] S. Du and M. Vuskovic, "Temporal vs. spectral approach to feature extraction from prehensile emg signals," in *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, 2004. IRI 2004.*, pp. 344–350, IEEE, 2004.

[14] P. Kiran, N. Abhiram, S. Taran, and V. Bajaj, "Tqwt based features for classification of als and healthy emg signals," *Am. J. Comput. Sci. Inf. Technol*, vol. 6, no. 2, p. 19, 2018.

[15] A. Phinyomark, C. Limsakul, and P. Phukpattaranont, "A novel feature extraction for robust emg pattern recognition," *arXiv preprint arXiv:0912.3973*, 2009.

[16] Z. Li, X. Li, Y. Liu, H. Xie, J. Li, F.-l. Wang, Q. Li, and X. Zhong, "Label supervised llama finetuning," *arXiv preprint arXiv:2310.01208*, 2023.

[17] T. Gao, J. Jin, Z. T. Ke, and G. Moryoussef, "A comparison of deepseek and other llms," *arXiv preprint arXiv:2502.03688*, 2025.

[18] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

Mitigating Malicious Spoofing and Sniffing Attacks Through a Portable Secure Router

1st Andrej Vuchkovikj
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
andrejvuchkovikj@outlook.com

2nd Viktor Danev
Business & IT
University of New York in Prague
Prague, Czech Republic
vddanev2004@gmail.com

3rd Viktor Denkovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
viktor.denkovski@uacs.edu.mk

4th Bogdan Jeliskoski
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
bogdan.jeliskoski@uacs.edu.mk

5th Goce Gavrilov
School of Computer Science and Information Technology
University American College Skopje
Skopje, R. N. Macedonia
goce.gavrilov@uacs.edu.mk

***Abstract*— Users face substantial security risks from the growing number of public Wi-Fi networks due to attacks such as spoofing and sniffing, compromising data privacy and security. The paper describes the development process of a portable security router built with a combination of Raspberry Pi hardware and OpenWRT software to protect against security risks. The Raspberry Pi takes advantage of its small design and processing power to function as an encrypted relay between users and threatened networks, defending against known internet hazards. OpenWRT's configuration combines with an SSH-type connection setup to form the methodology, alongside the integration of Wi-Fi adapters, which enable private network transmission. The network delivers stability through its essential configuration features regarding IP address assignment and DNS optimization. The results show that the device effectively establishes a secure public network connection for private Wi-Fi signal transmission to users. The document briefs methods for future improvement, such as VPN system integration, which would enhance security for users looking to delve deeper into internet privacy. Open-network security now has a low-cost and portable solution proving functional through this study, ultimately delivering benefits to travelling and security-wise users.**

Keywords*— Portable Router, Raspberry Pi, OpenWRT, Network Security, VPN

I. Introduction

Internationalization means an international threat. Public Wi-Fi is much more conveniently available across the globe. Access points are appearing more and more in short-term rentals, not just hotels and restaurants. Even if customers pay for the digital real estate for which they connect, many access points are left open or poorly secured, rendering anyone vulnerable to a spoofing or sniffing attack [1]. Spoofing is when an intruder sets up a false access point. Sniffing is when an intruder connects to a network and accesses the data packets traversing the network to obtain personally identifiable information [2]. In terms of the best hardware to use as a pocket-secure router, there is no better option than a Raspberry Pi. It's small, programmable, and has its hardware functionalities. Ultimately, this single-board computer can function as a secure local access point, with all connections running through it acting as a virtual barrier between the safe user and the suspicious public access network [3]. With OpenWRT, a customized Linux-based operating system (OS) made to route one's network, users can get internet access wherever they need it [4].

This paper covers the design and implementation of a portable secure router using a Raspberry Pi, supplemented by the OpenWRT application. Included are the command scripts, installation guidance, required hardware and software, purpose and troubleshooting of the device, which all culminate in an on-the-go device for better security. Additionally, recommendations for future improvements are provided, including the integration of a VPN for more secure use. Overall, this was a successful project and highly feasible for many people who have a plan on solidifying their data privacy when connecting to local public networks [5].

This paper is structured as follows: Section one introduces the research project. An explanation of the means and materials used to construct the portable router is provided in section two. Section three presents the results of the construction and testing. Section four contains a discussion of the results and suggestions for enhancements. Finally, section five provides a conclusion with the discoveries of this research and suggestions for further research within portable cybersecurity devices.

II. Methodology

For this research, OpenWrt version 24.10.0 was used as a specialized Linux OS targeting embedded devices [6]. It is a complete replacement for the vendor-supplied firmware of a wide range of wireless routers and non-network devices. The gain from a setup like this is not to create a single, static firmware, but to elevate the control one would have over their network. OpenWrt provides a fully writable filesystem with optional package management. This frees the user from restrictions of the application selection and configuration provided by the vendor and allows them to use packages to customize an embedded device to suit any application.

To implement such a system, a suitable platform was required. For this research, a Raspberry Pi was chosen to serve as a portable and secure router. Raspberry Pi is a series of small single-board-computers (SBCs); more specifically, a Raspberry Pi 4 model B was used to conduct this research [7].

Model B included all the necessary features for this research, such as:

- 1.5 GHz 64-bit quad-core ARM Cortex-A72 processor;

- On-board 802.11ac Wi-F; and
- Full Gigabit Ethernet.

This makes the Raspberry Pi 4 Model B an ideal choice for network testing and security research. The ability to handle wired and wireless networks efficiently made it particularly well-suited for evaluating OpenWrt's performance and potential application, which is convenient and used for testing future implementations.

The way the OS was integrated onto the Raspberry Pi was through its very convenient SD card slot. The SD card slot is used to store the OS, which is going to be running on the Raspberry Pi, as well as act as a storage device. The imager was used for installing the OS, and it is an official tool designed to simplify the process of flashing the OS onto a microSD card. This tool was selected due to its ease of use and compatibility with the Raspberry Pi hardware. By selecting the OpenWrt image within the software, the system is prepared efficiently, ensuring a streamlined and error-free installation process. The microSD card slot of the Raspberry Pi allowed for a quick integration, providing a simple yet effective way to deploy the custom firmware onto the device. Additionally, since the idea of the research was to create a portable secure router, another Wi-Fi card was necessary to be able to emit a private Wi-Fi network. For this purpose, the Alfa AWUSO36NHA wireless adapter was used. This high-gain Wi-Fi adapter is well known for its strong signal reception and compatibility with network security tools, making it an excellent choice for network auditing [8]. Equipped with the Atheros AR9271 chipset, which is widely supported for many Linux distros, making the driver installation simpler. The Alfa adapter utilizes a 2.4 GHz, one of the default frequencies that Wi-Fi devices operate on, as well as ensuring a stable connection between network devices. The adapter is connected to one of the four USB ports installed on the Raspberry Pi, ensuring reliable wireless communication even in scenarios requiring extended range or specialized configurations.

III. RESULTS

A. Initial Steps

The first initial step of the setup was to install the OS, which can be done by running the Raspberry Pi Imager and flashing the OpenWRT configuration on the microSD card. After deploying the image to the SD card, it can be inserted into the Pi itself. The following step was crucial to make this project work: the network setup and connection to the pi using SSH. This step was done and set up on a Windows OS. We unplug the Ethernet cable from the desktop or laptop. If it has an Ethernet port, one side is plugged into the Raspberry Pi and the other into the PC port. Then, navigate to the control panel and look for the section "Network and Internet" and select "View network status and tasks" Once here, the port on which the Pi and PC are interconnected is selected. After the section has been located, navigation to the IPv4 protocol was where a static assignment of the IP addresses is made. The PC uses an IP address of the format 192.168.1.1, a subnet of 255.255.255.0 and the most important IP would be the default gateway of the Pi which was 192.168.1.1. After this is done, a proper connection between the Pi and the PC needs to be confirmed. Now, the Raspberry Pi needs to be powered on, and access it via SSH. If everything is done correctly, the log-in page of OpenWRT should appear.

Following basic security protocol, the next crucial, suggestively mandatory step was to change the password of the new router, which can be done by simply entering "passwd" into the terminal, prompting the user to create a new password just like in any other Linux machine. Unlike usual Linux machines, to enter and configure, the user needs to open files using the "vi" (or "vim"), which is a text editor preinstalled on OpenWRT.

Amongst many configurations, the first requirement was to make a configuration on the "network" file. The network document is in the file path "/etc/config", making sure that not lose any of the progress made. It is highly advisable to make a copy of all the config files in order if something ends up being lost or misconfigured; as such, making sure that a backup is created is necessary. This is simply done by entering the command "cp network network.bk", and like that, the backup was created. This is done for all the config files located in the config directory. After the backup process, the next step is the configuration of all the network protocols.

By entering the network text document, the first configuration change would be the "option ipaddr 192.168.1.1" to set a new IP address and overwrite the default one. In this instance, the IP addresses set to static will be "10.28.10.1" This was the IP address that was decided to be set in this case, it can be any desired IP, as well instead of the default address of 192.168.1.1, this would be the address that from now on everyone connects to using SSH. Below the "option ip6assign" section, a new line must be added reading " option force_link '1' ", essentially prompting the device to force the newly set IP address as the default, whether the Pi is active or not. All of these changes need to be made in the section of "config interface 'lan'". To proceed further within the network file, a different interface should be generated titled "config interface 'wwan' ".

Afterwards, the user may add the following configurations:

- option proto 'dhcp' - This is telling the WAN interface to pull an IP address from DHCP from whichever network was trying to connect. An example of this would be public Wi-Fi at a coffee shop;
- option peerdns '0;
- option dns '1.1.1.1 8.8.8.8' - This utilizes which DNS server provider would be used. In this case the selected ones were Cloudflare and Google since they are considered one of the most stable options amongst many others.

To save the file: press Esc and type ":" followed by "wq" and finally press Enter. The newly written configuration allows the router to pull IP addresses through DHCP from any situational wireless network. Finally proceeding onto granting an internet connection, the router needs to be rebooted to load the newly set IP, which can be done by entering "reboot" in the terminal.

B. Connecting Raspberry Pi to the Internet

To configure the Raspberry Pi router, the user should modify the wireless settings file first. This is done by accessing the configuration file using the command vi

"/etc/config/wireless". Within, adjustments are made to the radio0 interface to ensure functionality (Fig. 1). Specifically:

- Set the option channel to 7. Add option hwmode '11g' beneath the existing option htmode line to enable 802.11g mode for improved performance;
- Change option disabled '1' to option enabled '1' or option disabled '0' to activate the interface;
- Create a new line: option short_gi_40 '0', to disable short guard intervals, which will benefit stability.

The following commands should be used to ensure the new changes are applied:

- uci commit wireless;
- wifi.

```
config interface 'loopback'
        option device 'lo'
        option proto 'static'
        option ipaddr '127.0.0.1'
        option netmask '255.0.0.0'

config globals 'globals'
        option ula_prefix 'fd5f:85ab:5427::/48'

config device
        option name 'br-lan'
        option type 'bridge'
        list ports 'eth0'

config interface 'lan'
        option device 'br-lan'
        option proto 'static'
        option ipaddr '10.7.7.1'
        option netmask '255.255.255.0'
        option ip6assign '60'
        option force_link '1'

config interface 'wwan'
        option proto 'dhcp'
        option peerdns '0'
        option dns '1.1.1.1 8.8.8.8'
```

Fig. 1. All the needed configurations needed to be changed to make a successful Static IP and DHCP transmission

C. Establishing Network Connectivity

The most effective way to get through this segment of the project is by using OpenWRT's graphical user interface (GUI) (Fig. 1). To access it, simply enter the Raspberry Pi's IP address into a preferred web browser. After logging in with the set password, navigate to /cgi-bin/luci/admin/network/wireless, where the newly configured radio user interface will be displayed (Fig. 3).

In this UI:

1. The "Scan" option will detect available networks;
2. Select the preferred network and enable "Replace wireless configuration" via its checkbox;
3. Enter the network's WPA passphrase in the designated field;
4. "Save and Apply" will confirm these settings.

The router has now begun receiving Internet access, which can be verified by executing a ping command in the terminal or by connecting the router to a new device via LAN cable.

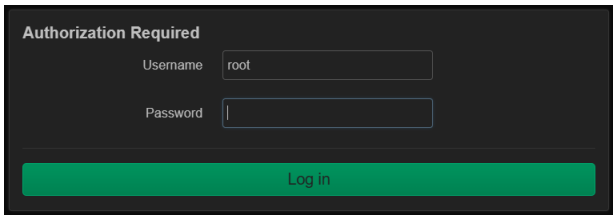


Fig. 2. GUI admin login page



Fig. 3. Home page of the GUI for OpenWRT, where all the connections happen

D. Setting Up the Wireless Adapter

After finalizing the previous step and connecting the Pi to a wireless network through its built-in wireless card, we can continue moving on and making this into an emitting router of that same network we have connected to. The first initial step is to run the following command: "opkg update" to update the list of packages that come installed with OpenWrt. There is one error that can come up in this process, and that is that the update might fail due to a network conflict. The easiest way to check and potentially solve this issue is to just reboot the Pi itself. The following step is accomplished by running "sudo reboot".

After the Pi boots up, the same command is typed again to see if the packages will get updated. If yes, then the next step is one of the more crucial ones to get the wireless card/interface to operate with the Pi. That would be installing the drivers. The following command works for most wireless cards, but it's good practice to check on the official website of OpenWrt whether the interface is compatible. After ensuring the interface is compatible, the following command is provided "opkg install kmod-rt2800-lib kmod-rt2800-usb kmod-rt2x00-lib kmod-rt2x00-usb kmod-usb-core kmod-usb-uhci kmod-usb-ohci kmod-usb2 usbutils openvpn-openssl luci-app-openvpn nano".

After successful installation, the next step is to check the USB list on the Pi; the command for this is "lsusb". This

command lists all the connected USB devices on the Pi, if the command is run before the adapter is plugged in, it won't be shown. This is done to detect the correct device once it is connected. After checking the devices and knowing what is currently active, we can then plug in our network interface and run the same command to check if it has successfully been detected.

The last step is to make the network interface active. This is achieved by entering the following command: ifconfig (name of wireless adapter) up. The name was left blank since this is a relative installation to the installation. Usually, it is supposed to be wlan1 for the additional adapter and wlan0 for the local adapter, which is preinstalled on the Pi. In the testing cases, that was not what happened. The name was different from what was expected, but that does not differ from the command execution once. The interface, wlan0, connects to a public Wi-Fi or any other Wi-Fi network. That in some terms would be the receiver since the wlan0 receives network connection, while the other network adapter transmits that same network or acts as an access point for the end user. To accomplish the transmitting protocol, what is needed is to edit the wireless network configuration file. To do this, the following command is used: “nano /etc/config/wireless”. This time, a different text editor was used, and that is the nano text editor, a much simpler text editor suited for this kind of editing. After opening the configuration file, a few configs will be shown, a few for the Pi interface and two more for the add interface card, one is going to be for the device config, and the other one will be for the wireless interface. The first thing that needs to be changed is to enable the device. This is accomplished by editing the line “option disabled ‘1’” into “option disabled ‘0”.

The next step would be to configure the desired SSID, which stands for the public network name people usually see when they want to connect to a Wi-Fi network. This is good to know and can detect our access point. Since the aim is to make a secure portable router, adding encryption is crucial. This is done by adding the following two lines in the wireless interface config:

1. “option encryption ‘psk2’”; and
2. “option key ‘(desired Wi-Fi password)’”.

After those configurations are done, proceed to save the file by pressing CTRL-X, Y and Enter; that should save the changes. Lastly, to commit the configuration, the following command is entered “uci commit wireless” and then “wifi” Now, it can be connected to the Raspberry Pi via wireless from any device to the specified network (Fig. 4).

```
config wifi-device 'radio1'
        option type 'mac80211'
        option path 'scb/fd500000.pcie/pci0000:00/0000:00:00.0/0000:01:00.0/usb1/1-1/1-1.4/1-1.4:1.0'
        option hwmode '11g'
        option channel '6'
        option htmode 'HT20'
        option disabled '0'
        option cell_density '0'

config wifi-iface 'default_radio1'
        option device 'radio1'
        option network 'lan'
        option mode 'ap'
        option ssid 'SucCesS'
        option encryption 'psk2'
```

Fig. 4. Explanation of the setting needed to be changed to enable the wireless adapter to transmit

E. System Usage Outcome Results

Finally, having the above full procedure done, the successful outcome generated results of the system usage can be seen in Fig. 5 and Fig. 6, The work is presented on the admin pages, where established passthrough connection on public Wi-Fi (Fig. 5), and Wi-Fi scaning on the public networks is being done by using wireless Pi interface (Fig. 6). As such, a new private network transmission is being generated for the end user.

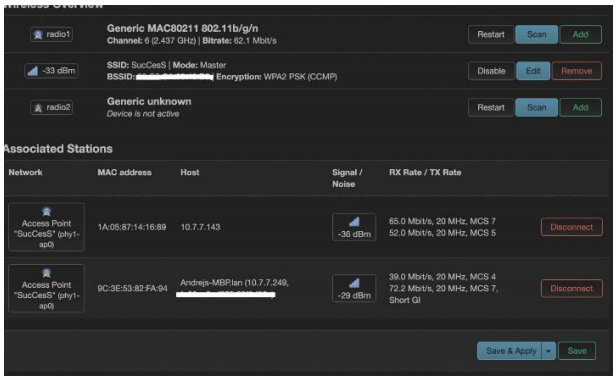


Fig. 5. Established a passthrough connection on public Wi-Fi

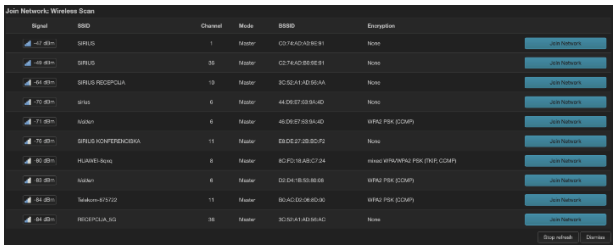


Fig. 6. Wi-Fi scan using the wireless Pi interface

IV. Discussion

A variety of malicious attacks occur daily. Not only are they a rising threat in popularity [9], but not much is done by the rest of the public to secure easily and freely accessible networks. The question asked here really translates to, should everyone have a pocket-sized secure router? Of course, the answer is both affirmative and negative. Personal data and information are shared and viewed by third-party sources to an extraordinary amount per day due to activities commonly shared by humans, including social media activity, comments, text messages, pictures exchanged, reactions, etc [10]. Therefore, it depends on how much an individual values their virtual privacy. In the sense of security, the general public should invest in a device much like this one. It is an undeniable fact that utilizing a secure router while travelling will increase browsing and data privacy by enabling the comfort of the eradication of man-in-the-middle vulnerabilities [11]. Though there are notable methods that could improve the Pi router’s security, namely by integrating a virtual private network (VPN).

V. Conclusion

This study shows that creating a portable secure router is achievable and essential for the current digitally connected world we live in. As international travel increases and more public Wi-Fi networks are created, from hotels to restaurants to Airbnb, experiencing hacking from spoofing and sniffing attack vectors is a troubling and harsh reality. Therefore, the portable secure router, based on Raspberry Pi hardware and OpenWRT software, is something one can create to sidestep this vulnerability and create a barrier between the user and public access networks that could otherwise lead to dangerous situations.

This project showcased the capability and potential for building a portable secure router. Public Wi-Fi access is everywhere, meaning the more people access public Wi-Fi, the more likely spoofing and sniffing attacks will be on the rise. Yet with a Raspberry Pi and OpenWRT, merely constructing a portable secure router, such a device will ease worry, as it provides a firewall between the user's device and the public malicious network.

The results presented imply that anyone can achieve enhanced cybersecurity with instantly available hardware and open-source software, and minimal compromise on portability. In addition, the specific size, configurability, and hardware features of the Raspberry Pi make it an ideal device for the task, allowing anyone to connect securely from anywhere. This project shows that with low-cost components and instantly available open-source software, anyone can strengthen their security while travelling. In addition, the Raspberry Pi's versatility and ability to integrate hardware expansions make it a perfect device for a project that anyone can use to manage their network access while in a compromised position.

The fact that something like a mobile secure router even exists shows how much safety is still desired in a world where hacking is both advanced and far too common. Thus, while testing, this router will act as the stable connection proxy between what someone would need to connect to in a public space and their private devices, minimizing the chances of hacking, stolen information, or privacy invasion where such a violation would otherwise be expected.

Future work needs to evaluate if further protective options can be implemented, like VPN access, IDS, and security-related automatic software updates. Furthermore, improving the ease of use and making the wearable non-complex for non-technical users would be a big step forward to general applicability. The portable secure router is the ultimate digital protection device on the go, and when using public Wi-Fi. Finally, as technology continues to grow and advance worldwide, increasingly portable devices like the one created for this project will be a necessity for personal digital protection.

References

[1] Sasi, T., Lashkari, A. H., Lu, R., Xiong, P., and Iqbal, S. "A Comprehensive Survey on IoT Attacks: Taxonomy, Detection Mechanisms and Challenges." *Journal of Information and Intelligence*, vol. 2, no. 6, Elsevier BV, Dec. 2023, https://doi.org/10.1016/j.jiixd.2023.12.001.

[2] None Saddam RA, None Angga Pranata, None Sugiono, None Rizki Zulanggara, None Nur Halimah, Nur, S., None Rosdiana SM, None Nurhalim, and None Aprina Handayani. "Sniffing and Spoofing in Computer Security." *International Journal of Integrative Sciences*, vol. 2, no. 6, June 2023, pp. 881–86, https://doi.org/10.55927/ijis.v2i6.4528. Accessed 10 Oct. 2024.

[3] Osman, M. N., Sedek, K. A., Othman, N. A., Rosli, M. A., and Maghribi, M. "Enhancing Security and Privacy in Local Area Network (LAN) with TORVPN Using Raspberry Pi as Access Point: A Design and Implementation." *Journal of Computing Research and Innovation*, vol. 6, no. 2, Sept. 2021, pp. 29–40, https://doi.org/10.24191/jcrinn.v6i2.190. Accessed 9 Dec. 2021.

[4] Palazzi, C. E., Matteo Brunati, and Roccetti, M. "An OpenWRT Solution for Future Wireless Homes." *CiteSeer X (the Pennsylvania State University)*, Pennsylvania State University, July 2010, https://doi.org/10.1109/icme.2010.5583223.

[5] Vasconcelos, M., Vasconcelos, M., Ellabidy, M., Albanese, D., and Yun, M. PiShield: Low-Cost Shield for Public Wi-Fi. cs.bc.edu/~yunmd/pdfs/HICE2023_PS.pdf. Accessed 1 Apr. 2025.

[6] "[OpenWrt Wiki] OpenWrt 24.10.0 - First Stable Release - 6. February 2025." Openwrt.org, 2025, openwrt.org/releases/24.10/notes-24.10.0. Accessed 1 Apr. 2025.

[7] Ltd, Raspberry Pi (Trading). "Buy a Raspberry Pi 4 Model B." Raspberry Pi, 2023, www.raspberrypi.com/products/raspberry-pi-4-model-b/.

[8] "AWUS036NHA." ALFA Network Inc., www.alfa.com.tw/products/awus036nha?variant=3647396616608.

[9] Vuchkovikj, A., Kostadinov, N., Denkovski, V. and Gavrilov, G. "Accessing a Windows Machine Using a Makeshift Rubber Ducky." CIIT 2024, 19 Apr. 2024, https://ciit.finki.ukim.mk/proceedings/ciit-2024-proceedings.pdf. Accessed 1 Apr. 2025.

[10] Sangeen, M., Bhatti, N. A., Kifayat, K., Alsadhan, A. A., and Wang, H. "Blind-Trust: Raising Awareness of the Dangers of Using Unsecured Public Wi-Fi Networks." Computer Communications, vol. 209, Sept. 2023, pp. 359–67, https://doi.org/10.1016/j.comcom.2023.07.011.

[11] Obaidat, M. A., Obeidat, S., Holst, J., Al Hayajneh, A., and Brown, J. "A Comprehensive and Systematic Survey on the Internet of Things: Security and Privacy Challenges, Security Frameworks, Enabling Technologies, Threats, Vulnerabilities and Countermeasures." Computers, vol. 9, no. 2, May 2020, p. 44, https://doi.org/10.3390/computers9020044.

Ollama LLMs serverless deployment - Performance evaluation using LangChain

Zorica Karapancheva
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
zorica.karapancheva@finki.ukim.mk

Jovan Simonoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
jovan.simonoski@students.finki.ukim.mk

Marjan Gushev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
marjan.gushev@finki.ukim.mk

Dimitar Mileski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
dimitar.mileski@finki.ukim.mk

***Abstract*—The emergence of Open Source LLMs has created opportunities for deployment in various environments, spanning from the edge to the cloud across the entire cloud continuum. This paper presents a scalability performance evaluation of 7 different Ollama LLMs in a serverless environment. Scalability Evaluation metrics such as Time To First Token (TTFT), Time Per Output Token (TPOT), End-to-End Request Latency, and Throughput were measured using LangSmith's monitoring services. The results were then analyzed in terms of response quality, using evaluation frameworks and metrics such as summarization, answer relevance, faithfulness, contextual recall, and precision. However, this paper does not conduct an independent evaluation of response quality; instead, it references the findings from the models' official papers and documentation.**

***Index Terms*—large language models, scalability, serverless, Ollama**

I. Introduction

The rapid advancement of Large Language Models (LLMs) has led to a growing interest in their deployment across diverse computing environments, from edge devices to large-scale cloud infrastructures. While proprietary LLMs have dominated many applications, the emergence of open-source alternatives has introduced new opportunities for customization, accessibility, and cost-effective scaling. Among these, Ollama LLMs have gained traction due to their adaptability and ease of deployment.

Scalability remains crucial in assessing the feasibility of deploying LLMs in real-world applications, particularly in serverless environments, where resource allocation and execution dynamics differ significantly from traditional computing paradigms. To address this, our study conducts a comprehensive scalability performance evaluation of 7 different Ollama LLMs in a serverless setting. We assess key metrics, including Time To First Token (TTFT), Time Per Output Token (TPOT), End-to-End Request Latency, and Throughput, utilizing LangSmith's monitoring services for precise measurement.

Beyond raw performance, response quality is an essential consideration when deploying LLMs. While this paper does not independently evaluate the response quality of the tested models, it contextualizes performance results using established evaluation frameworks. Metrics such as summarization accuracy, answer relevancy, faithfulness, contextual recall, and precision are referenced from each model's official papers and documentation.

By combining performance analysis with qualitative insights, this study provides valuable guidance for researchers and practitioners seeking to optimize open-source LLM deployments in serverless environments. The findings highlight the trade-offs between scalability and response quality, offering a foundation for informed decision-making in selecting LLMs for various real-world applications.

RQ1 Which Open Source Ollama LLMs scale the best (measured by TTFT, TPOT, end-to-end request latency, and throughput) in a serverless environment?

RQ2 How does this performance relate to response quality as measured by MMLU, MATH, and HumanEval evaluation frameworks?

The structure of the paper is as follows: Section II analyzes the similarities and differences in the related work. Methods in Section III specify the solution architecture and workflow, experiments, and the evaluation metrics to evaluate the results. The results presented in Section IV are assessed and discussed in Section V. Section VI presents the conclusion and future work.

II. Related Work

The rapid integration of Large Language Models (LLMs) into various applications has prompted extensive research into their performance, scalability, and deployment efficiency. Prior studies have analyzed the capabilities of LLMs across multiple tasks, including document retrieval, summarization, code generation, chatbot applications, and reasoning, with a focus on accuracy, latency, and computational efficiency.

Ko et al. [1] demonstrated that Llama 3.1 improves responses in retrieval-augmented generation (RAG) tasks, while Phi3 offers lower latency. Llama3, Gemma, and Mistral achieved competitive ROUGE scores when summarizing CNN and Daily Mail articles, suggesting that Llama3 performs comparably to existing methods in this domain [2].

However, LLMs failed to achieve reasonable scores in code generation, indicating that current models still struggle to generate accurate code based on ground truths. Code-specific models consistently outperformed general-purpose models, reinforcing the need for specialized architectures in software development [2].

The chatbot use case presents a different challenge - models like Llama3, Gemma, and Mistral produced reasonable lexical similarity scores. Still, evaluation metrics remain problematic, as responses may be factually correct yet lexically different from the reference ground truth, leading to artificially lower scores [2]. Similarly, LLAVA models, which integrate vision and language processing, showed reduced accuracy compared to text-only models of similar size. For example, llava-llama3:8b-v1.1-fp16 achieved only 47.34% accuracy in website categorization, lagging behind comparable text-only models [3]. These findings highlight the trade-offs inherent in multimodal model design and underscore the need for more effective evaluation strategies.

As LLMs scale, they introduce challenges related to computational cost, memory footprint, and inference efficiency. Traditional training and deployment methods become expensive, necessitating the development of optimized architectures and serving strategies [4]. Ollama provides a framework for running open-source LLMs efficiently on local machines [5] and cloud-based environments [6]. Studies evaluating Ollama's deployment scenarios have explored local execution, multi-GPU distributed inference, and cloud-based serverless deployments, each with different cost, latency, and resource utilization trade-offs. Watts et al. analyzed scalability trade-offs in cloud-hosted LLMs, comparing inference latency across different model sizes and deployment setups [6]. Furthermore, some models, such as "gemma:7b-instruct-q8_0," exhibited high GPU memory usage relative to their parameter count, suggesting room for optimization in hardware-aware deployment strategies [3]. Liu et al. examined the performance of locally deployed Ollama models in Retrieval-Augmented Generation (RAG) tasks, highlighting how optimizing retrieval mechanisms can significantly improve efficiency in constrained computational environments [7].

LangChain has emerged as a robust framework for integrating LLMs into applications, offering features such as memory management, RAG, and optimized API calls. LangChain has proven to be a valuable tool for LLM-based development, introducing structured frameworks and reusable components that enhance flexibility and interoperability with multiple data sources [8], [9]. While prior research has demonstrated that LangChain improves query efficiency in structured retrieval tasks [10], there has been limited exploration of its impact on serverless LLM performance at scale.

III. Methods

A. Testing Architecture

We deployed each model using Google Cloud Run, a managed serverless platform, as illustrated in Fig. 1, to assess the scalability and performance of 7 different Ollama LLMs in a serverless environment.

Each model instance was provisioned with eight vCPUs, 16 GiB of memory, and a maximum of 50 instances to handle concurrent requests efficiently. The deployment was

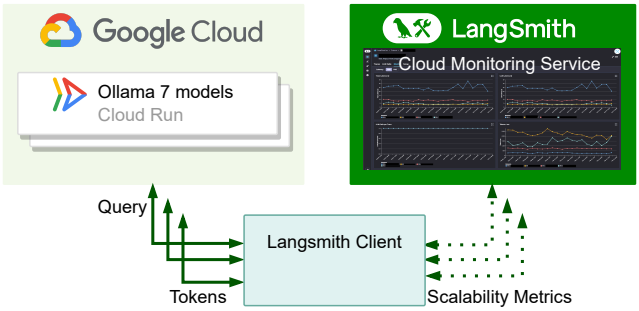


Fig. 1: Evaluation Architecture

configured with four concurrent requests per instance (-concurrency 4) and parallel execution of up to four model inferences (OLLAMA_NUM_PARALLEL=4). CPU throttling was disabled (–no-cpu-throttling), and a 600-second timeout was set to ensure the completion of longer-running requests.

B. Experiments

The research questions RQ1 and RQ2 evaluate the relationship between scalability metrics and response quality. The scalability metrics analyzed include TTFT, TPOT, End-to-End Request Latency, and Throughput. The MMLU, MATH, and HumanEval evaluation frameworks are used to assess the quality of responses.

- **Model selection** We selected seven open-source Ollama LLMs with similar sizes—7B, 8B, and 9B parameters—ensuring optimal deployment on 16GB instances. The evaluated models are: dolphin3-8b, granite-code-8b, hermes3-8b, mistral-openorca-7b, orca-mini-7b, starcoder-7b, and yi-9b. The models were deployed in a serverless environment using Cloud Run for their evaluation.
- **Experiment execution** Each model is tested with three questions, one from each evaluation framework: MMLU, MATH, and HumanEval. The cold start phenomenon in serverless environments has been widely explored in previous research; however, it is beyond the scope of this paper. To eliminate serverless cold starts, we first run the experiment once to avoid the cold start effect. Then, we measure the scalability evaluation metrics (TTFT, TPOT, latency, and throughput) for each question.

C. Evaluation Methodology

Measuring the performance of Large Language Models (LLMs) extends beyond metrics involving tokens per second (TPS), as tokenization strategies vary across different models. Some models generate complete words, others break them into sub-words, and some tokenize at the byte level [11]. Performance evaluations should instead focus on alternative benchmarks that evaluate the model's efficiency. To overcome these challenges, the evaluation methodology focuses on four key performance metrics (as shown in Fig. 2) - Time To First Token, Time Per Output Token, End-to-End Request Latency, and Throughput [12]. These metrics provide a complete overview of an LLM's efficiency and responsiveness.

- **TTFT** represents the time elapsed between a user query submission and the model generating the first output

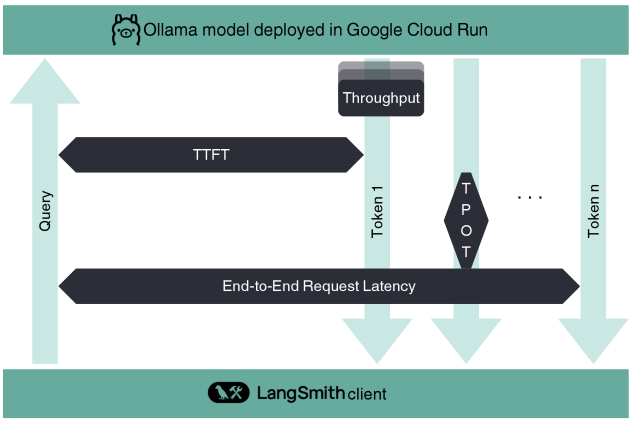


Fig. 2: Scalability Evaluation metrics

token. This metric enhances real-time interactions by allowing users to review responses before the output is fully generated.

- **TPOT** represents the time required to generate each additional token after the initial one, directly affecting perceived responsiveness and user experience. This metric serves as a valuable benchmark for the efficiency of LLMs in extended text generation, facilitating natural interactions.
- **End-to-End Request Latency** represents the total time from submitting the query to receiving the entire response. This includes network latency, batching, queuing delays, and the model's processing time. In general, it is equal to the sum of TTFT and the response generation time.
- **Throughput** represents the total number of tokens per second an inference server can generate through all active user requests. This metric reflects the overall efficiency and scalability of the system.

In addition to efficiency, model accuracy is crucial in evaluating the performance of LLMs. To provide a broader perspective, for all of the models in the study, we reference benchmark results from evaluation datasets, including MMLU (Massive Multitask Language Understanding) for general knowledge and reasoning [13], MATH for mathematical problem-solving [14], and HumanEval for code generation capabilities [15]. These benchmarks offer consistent evaluations of model accuracy across various fields. Although the accuracy scores for the evaluated models are not measured in this study, they are compiled from existing literature and presented in Table 1. By highlighting potential variations in model benchmark results across tasks, this comparison offers valuable insight into how LLMs scale in serverless environments, particularly on benchmarks such as MATH, HumanEval, and MMLU.

IV. Results

Fig. 3, 4, and 5 present the evaluation metrics for MMLU, MATH, and HumanEval, respectively. Each figure illustrates the scalability evaluation of the 7 Open-Source Ollama LLMs deployed in a Cloud Run serverless environment.

For each plot, the y-axis represents the value of the scalability metrics (such as TTFT, TPOT, latency, and throughput), while the x-axis shows the various Ollama models. These metrics offer insights into how each model scales in a serverless environment and its corresponding impact on overall performance, as evaluated by the MMLU, MATH, and HumanEval frameworks.

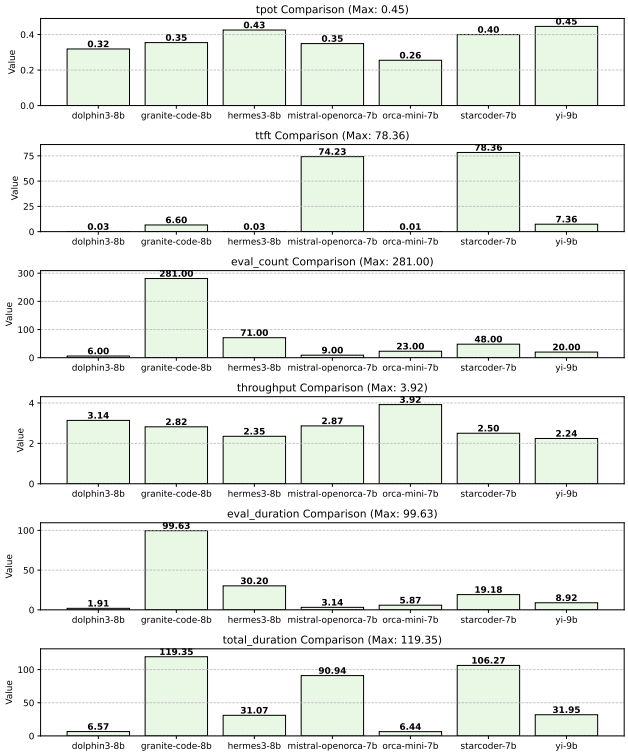


Fig. 3: MMLU

By analyzing these plots, we can observe trends in how scalability affects model performance across different models and how Ollama models compare against each other in terms of both scalability and response quality.

V. Discussion

Two primary research questions guide this research. RQ1 investigates which of the seven Open Source Ollama LLMs performs best in terms of scalability (TTFT, TPOT, End-to-end request latency, throughput) in a serverless environment, while RQ2 explores how the performance of these models correlates with their response quality (MMLU, MATH, HumanEval). In the following discussion, we examine the results of the scalability tests, providing insights into the trade-offs between scalability and response quality for each of the models under consideration.

In analyzing the results of the MMLU, MATH, and HumanEval benchmarks, we can observe distinct trends in how the scalability metrics (TPOT, TTFT, throughput, and end-to-end latency) correlate with model performance. Although the models' advantages and weaknesses vary depending on the metrics, certain leaders emerge in terms of response times and scalability.

Regarding **MMLU**, *Yi-9b* demonstrates the best overall performance in terms of scalability, achieving the highest TPOT score of 0.45 and maintaining a relatively low latency of 31.95 seconds. This model also performs well in terms

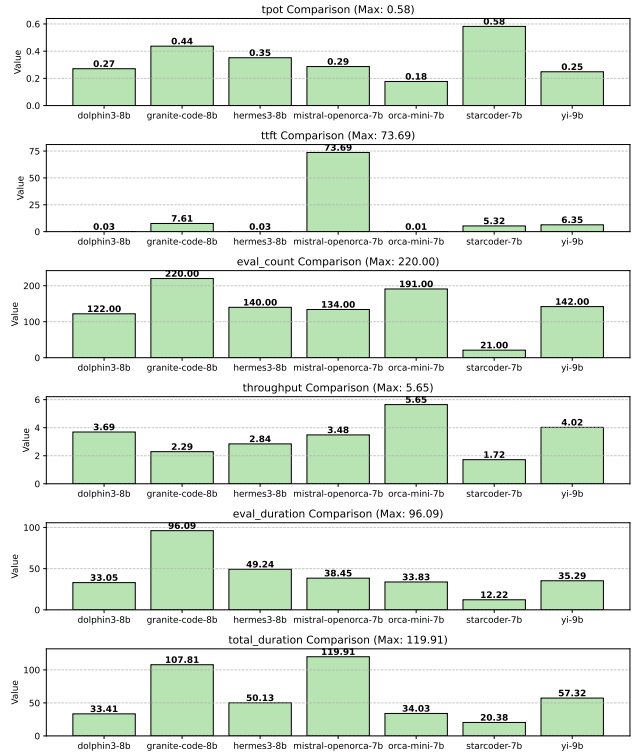


Fig. 4: MATH

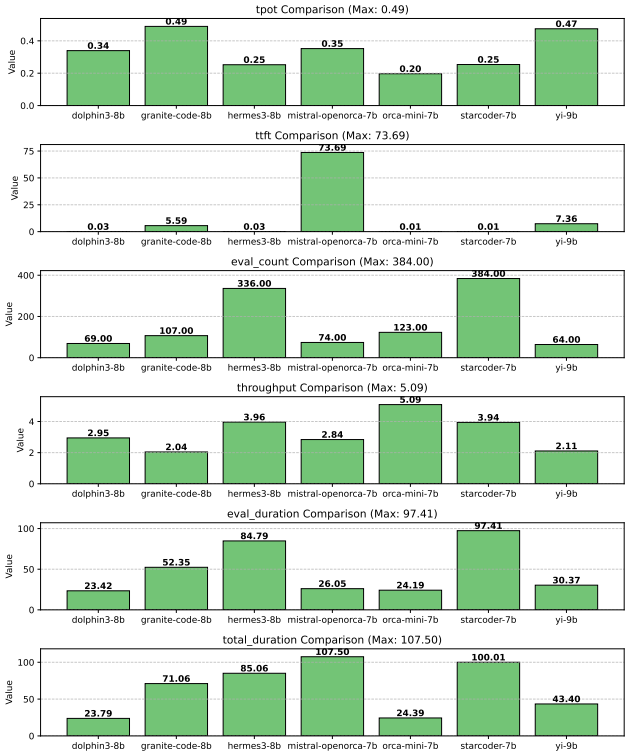


Fig. 5: HumanEval

of throughput. However, its moderate performance in TTFT (7.36) indicates that while it delivers quick responses overall, there is still room for improvement in the initial time-to-first-token. On the other hand, *Orca-Mini-7b* stands out in terms of scalability, with the lowest TTFT of 0.01 and the highest throughput of 3.92. While its TPOT score is lower at 0.26, this model presents outstanding response time and efficiency, making it a competitive option in serverless environments that prioritize speed. In contrast, models such as *Granite-Code-8b* and *Mistral-Openorca-7b* exhibit significantly higher latency and longer total durations, indicating that while they may perform adequately in certain use cases, they struggle to scale effectively in real-time applications. In terms of scalability compared to the accuracy of the models, it is notable that *Mistral-Openorca-7b*, which has the most prolonged latency, has the highest measured accuracy, as shown in other literature (Table 1). On the other hand, *Orca-Mini-7b*, even with great scalability, has poor accuracy ratings.

Regarding **MATH**, *Starcoder-7b* demonstrates intense competition with the highest TPOT of 0.58. This indicates its superior performance in scalability for mathematical problem-solving, although its latency of 20.38 seconds and TTFT of 5.32 highlight challenges in minimizing response times. Despite these trade-offs, *Starcoder-7b* remains one of the better performers in terms of scalability within this metric. *Yi-9b* again performs well, though its TPOT is slightly lower at 0.25, the model compensates with its competitive throughput and low latency, making it a balanced option overall. In contrast to *Granite-Code-8b* and *Mistral-Openorca-7b*, which also display poor scalability, with high latencies and TTFT values that limit their potential in real-world applications. Regardless of the scalability differences between *Starcoder-7b* and *Yi-9b*, they achieve similar accuracy results. Furthermore, *Mistral-Openorca-7b*, besides poor scalability, also does not perform well in terms of accuracy, with a MATH percentage of only 3.55.

Regarding **HumanEval**, *Yi-9b* again stands out as the best-performing model, delivering a TPOT of 0.47 and maintaining low latency, which is essential for real-time code generation tasks. Its throughput is somewhat lower than that of other models, such as *Orca-Mini-7b*. Still, the balance between response time and scalability makes it the top performer in the HumanEval benchmarking. *Orca-Mini-7b* excels in throughput and scalability with a TPOT of 0.20; however, its latency is still relatively high, which may impact its usability in more time-sensitive scenarios. *Granite-Code-8b* and *Starcoder-7b*, although decent TPOT scores of 0.49 and 0.25, respectively, are challenged by high latencies and low throughput, ultimately limiting their effectiveness in HumanEval tasks that require quick and consistent performance. In terms of HumanEval benchmarking, a lot of the models have missing accuracy readings. However, among the available options, it is notable that *Yi-9b* has the highest HumanEval accuracy, achieving 39.0%.

Overall, *Yi-9b* is the standout performer across MMLU and HumanEval, providing the best balance of scalability, throughput, and low latency. However, for specific use cases where response times and scalability in MATH tasks are more critical, *Starcoder-7b* and *Orca-Mini-7b* emerge as strong competitors, with *Starcoder-7b* leading in scalability but still struggling with latency and throughput. Meanwhile, models like *Mistral-Openorca-7b* and *Granite-Code-8b*. However,

providing reliable evaluation counts demonstrates significant limitations in scalability, particularly in terms of latency and response times, making them less viable in real-time, high-performance environments.

VI. Conclusion

This paper evaluated the scalability of seven open-source Ollama LLMs in a serverless environment, analyzing key performance metrics such as TTFT, TPOT, end-to-end request latency, and throughput. The results indicate that 'Yi-9b' consistently delivers the best balance between scalability and performance, excelling across MMLU and HumanEval benchmarks with strong TPOT scores, low latency, and competitive throughput. In MMLU, it achieves high accuracy while maintaining efficient scalability. HumanEval demonstrates strong response quality for code generation tasks. 'Orca-Mini-7b' emerges as the most efficient model in terms of response times, achieving the lowest TTFT and highest throughput, making it an optimal choice for speed-sensitive applications. However, its accuracy in MMLU is lower. 'Starcoder-7b', while demonstrating strong scalability in MATH tasks with the highest TPOT score, faces challenges with latency and throughput, though it maintains competitive accuracy in mathematical problem-solving. Models such as 'Granite-Code-8b' and 'Mistral-Openorca-7b' exhibit significant limitations in real-time performance due to high latencies and poor scalability. Additionally, 'Mistral-Openorca-7b' ranks high in MMLU accuracy but struggles in MATH and HumanEval. These findings underscore the importance of selecting models tailored to specific use case requirements, thereby striking a balance between response quality and computational efficiency in serverless environments.

References

[1] K. Ko, T. Y. Nyein, K. K. Oo, T. Z. Oo, and T. T. Zin, "Retrieval augmented generation for document query automation using open source llms," in *2024 5th International Conference on Advanced Information Technologies (ICAIT)*. IEEE, 2024, pp. 1–6.

[2] I. Ul Haq, "Performance analysis of large language models on summarization, code generation, and conversational ai," 2024.

[3] D. Sava, "Text-based classification of websites using self-hosted large language models: An accuracy and efficiency analysis," B.S. thesis, University of Twente, 2024.

[4] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, R. Sauvestre, T. Remez *et al.*, "Code llama: Open foundation models for code," *arXiv preprint arXiv:2308.12950*, 2023.

[5] T. Isachenko and S. Bhuiyan, *Generative AI with local LLM*. Timur Isachenko, 2024.

[6] I. Watts, V. Gumma, A. Yadavalli, V. Seshadri, M. Swaminathan, and S. Sitaram, "Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data," *arXiv preprint arXiv:2406.15053*, 2024.

[7] F. Liu, Z. Kang, and X. Han, "Optimizing rag techniques for automotive industry pdf chatbots: A case study with locally deployed ollama modelsoptimizing rag techniques based on locally deployed ollama modelsa case study with locally deployed ollama models," in *Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Intelligent Information Processing*, 2024, pp. 152–159.

[8] 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.

[9] B. Auffarth, *Generative AI with LangChain: Build large language model (LLM) apps with Python, ChatGPT, and other LLMs*. Packt Publishing Ltd, 2023.

[10] T. P. Jacob, B. L. S. Bizotto, and M. Sathiyanarayanan, "Constructing the chatgpt for pdf files with langchain–ai," in *2024 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2024, pp. 835–839.

[11] M. Agarwal, A. Qureshi, N. Sardana, L. Li, J. Quevedo, and D. Khudia, "Llm inference performance engineering: Best practices," 2024, accessed: 30 March 2025. [Online]. Available: https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices

[12] NVIDIA, "Llm benchmarking metrics guide," 2025, accessed: 30 March 2025. [Online]. Available: https://docs.nvidia.com/nim/benchmarking/llm/latest/metrics.html

[13] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[14] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset, 2021," *URL https://arxiv. org/abs/2103.03874*, 2024.

[15] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

TABLE I: Performance Benchmarks Across Models

Model	MMLU	MATH	HumanEval
deepseek-r1:7b	90.8[1]	97.3[1]	73.2[2]
mistral:7b	60.1[3]	13.1[3]	30.5[3]
qwen:7b	58.2[4]	11.6[4]	29.9[4]
qwen2:7b	70.5[1]	49.6[1]	79.9[1]
qwen2.5:7b	70.5[5]	-	-
gemma:7b	64.3[6]	24.3[6]	32.3[6]
gemma2:9b	71.3[1]	36.6[1]	40.2[1]
llama2:7b	45.3[7]	14.6[7]	29.9[7]
llama3:8b	68.4[1]	30.0[1]	62.2[1]
llama3.1:8b	73.0[1]	51.9[1]	72.6[1]
llava:7b	-	-	-
codellama:7b	31.12[8]	-	24.36[8]
starcoder2:7b	7.14[9]	3.1[9]	26.1[9]
codegemma:7b	56.59[10]	-	30.6[10]
openthinker:7b	35.16[11]	42.6[11]	-
minicpm-v:8b	-	-	-
llava-llama3:8b	-	-	-
wizardlm2:7b	42.7[12]	3.93[13]	19.1[12]
dolphin-mistral:7b	20.23[14]	5.06[14]	-
dolphin-llama3:8b	61.04[15]	5.74[15]	-
orca-mini:7b	29.97[16]	-	-
yi:9b	68.4[17]	15.9[17]	39.0[17]
hermes3:8b	64.79[18]	7.48[18]	-
doplhin3:8b	-	-	-
zephyr:7b	61.06[19]	2.87[19]	-
olmo2:7b	63.7[20]	31.3[20]	-
granite-code:8b	-	21.4[21]	-
starcoder:7b	33.9[22]	15.1[22]	33.6[22]
vicuna:7b	47.1[23]	1.36[13]	-
mistral-openorca:7b	62.24[24]	3.55[24]	35.9[25]

[1] https://ollama.com/library/{model_name}
[2] https://medium.com/@amareswer/deepseek-r1-the-next-evolution-in-ai-language-models-4f22832aa648
[3] https://mistral.ai/news/announcing-mistral-7b
[4] https://github.com/QwenLM/Qwen
[5] https://github.com/QwenLM/Qwen2.5
[6] https://ai.google.dev/gemma/docs/core/model_card
[7] https://arxiv.org/abs/2307.09288
[8] https://llm.extractum.io/model/codellama%2FCodeLlama-7b-hf,9wLn4CFDFFUu7QIbKH6D1
[9] https://llm.extractum.io/model/bigcode%2Fstarcoder2-7b,15DU1x3WFfv65tU16NRqbA
[10] https://llm.extractum.io/model/google%2Fcodegemma-7b,C55I3mhJwhGLkwLF55Blg
[11] https://llm.extractum.io/model/open-thoughts%2FOpenThinker-7B,5uh5EDODQOk4YbhrXeoOrZ
[12] https://github.com/nlpxucan/WizardLM
[13] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard
[14] https://llm.extractum.io/model/cognitivecomputations%2Fdolphin-2.9.3-mistral-7B-32k,7jO5C5xqU9DsOptOEq7UDY
[15] https://llm.extractum.io/model/cognitivecomputations%2Fdolphin-2.9-llama3-8b,46qrCwJiXHvjZEUZzVRXeH
[16] https://huggingface.co/pankajmathur/orca_mini_7b
[17] https://github.com/01-ai/Yi
[18] https://arxiv.org/pdf/2408.11857
[19] https://llm.extractum.io/model/HuggingFaceH4%2Fzephyr-7b-beta,7JLYCMBpBKlhK9cdxgn4eR
[20] https://allenai.org/blog/olmo2
[21] https://arxiv.org/pdf/2405.04324
[22] https://arxiv.org/pdf/2305.06161
[23] https://arxiv.org/pdf/2306.05685
[24] https://llm.extractum.io/model/Open-Orca%2FMistral-7B-OpenOrca,1GWIbwevXD7b4iNXQS9mcY
[25] https://huggingface.co/uukuguy/Mistral-7B-OpenOrca-lora-merged

Combination of Kolb's Experiential Learning Cycle and Project-Based Learning in Data Science Education

Nevena Kimova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
nevena.kimova@students.finki.ukim.mk

Prof. Ana Madevska Bogdanova, PhD
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
ana.madevska@finki.ukim.mk

***Abstract*—Today, teaching methods struggle to keep up with the latest developments and trends in the field. This paper identifies the weak connection between theory and practice and proposes a solution that has the potential to overcome the limitations of traditional teaching methods. The proposed solution is to integrate Kolb's experiential learning cycle into project-based learning while teaching data science. This creates an environment where students not only solve problems, but also learn to think scientifically. In this way, in addition to being able to connect theory with practice, they are also encouraged to think critically and be creative, which are essential skills for the field. The idea is to change perspectives and transform the way data science is taught and learned.**

I. Introduction

Despite strong foundations in machine learning, many graduates struggle to apply their knowledge to real problems. For instance, a machine learning specialist may struggle when confronted with messy, unstructured data, while a junior analyst lacks the judgment to select appropriate methods for a given problem. These scenarios highlight the disconnection [1] between conceptual understanding and practical problem-solving—a key challenge in data science education. Nowadays, organizations rely heavily on data scientists to derive insights. However, traditional approaches fail to prepare students with the necessary skills to meet these demands. While theoretical knowledge is essential, it is insufficient without competencies in critical thinking and adaptive problem-solving. Consequently, many professionals enter the workforce unprepared for the data challenges. Project-based learning offers [2] a framework to address this gap. It serves as an experiential learning tool that bridges abstract concepts with concrete applications. However, the completion of projects is not enough, but reflective practice is much needed. Students must evaluate their work by identifying failures, successes and learn from the outcomes. To support this, we propose integrating Kolb's experiential learning cycle [3] within PBL frameworks. Kolb's model, with its emphasis on concrete experience, reflective observation, abstract conceptualization, and active experimentation, accommodates diverse learning styles while promoting iterative skill development.

The urgency of reforming data science education cannot be overstated. As the field evolves, educators must prepare students for a complex and unpredictable professional environment. Traditional methods are inadequate for cultivating the dynamic, adaptive expertise required in modern data science. This paper explores an approach that combines PBL with a reflective learning cycle that can better align with the industry needs.

II. Kolb's Experimental Learning Cycle

David A. Kolb [4] developed the Experiential Learning Cycle to formalize how individuals learn through experience. This model shows that learning is an active, iterative process where knowledge is constructed through the transformation of experience, rather than through passive absorption.

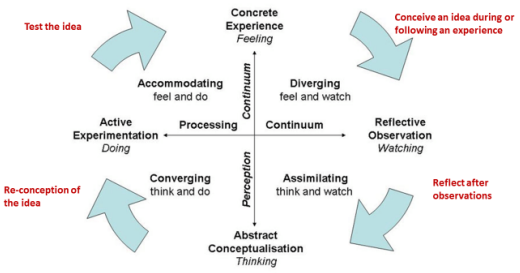


Fig. 1. Kolbs Learning Cycle and Learning Styles.

A. The Four Phases

Kolb's cycle [5] transforms experience into knowledge by structuring learning into four interdependent phases:

- **Concrete experience** - The learner engages in a direct, hands-on experience, either new or revisited one through a new conceptual lens.
- **Reflective observation** - The learner reflects on the experience, analyzing it in relation to prior knowledge.

- **Abstract conceptualization** - Reflection leads to forming new hypotheses or the refinement of existing theories.
- **Active experimentation** - The learner tests the concepts and generates experiences for iterative learning.

By integrating Kolb's cycle into curricula, educators can create learning environments that prioritize active participation, critical reflection, and adaptive problem-solving.

B. Project-based learning

Project-Based Learning(PBL) [6][7] is an instructional methodology where students acquire knowledge and skills by investigating and responding to complex problems over extended periods. Unlike traditional learning approaches, PBL emphasizes student engagement, interdisciplinary thinking, and real-world applicability.

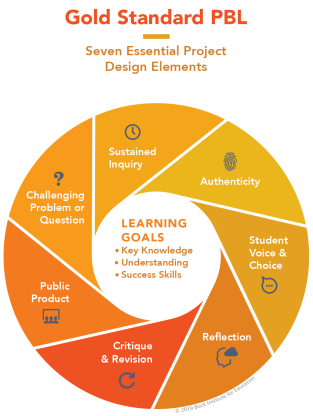


Fig. 2. Gold Standard PBL.

Steps in Project-Based Learning

- **Challenging Problem** - Real problem that aligns with learning objectives and connects to students' interests which makes learning meaningful.
- **Inquiry** - Students iteratively ask questions, research, refine solutions ny fostering curiosity and understanding.
- **Authenticity** - Projects imitate real-world contexts, tools, and constraints, bridging theory and practice.
- **Student Contribution** - the concepts are put into the experimental stage. Learners exercise control over project design, promoting ownership and leadership skills.
- **Reflection** - Various methods are used for structured reflection on what has been done and discuss the outcomes.
- **Revision** - Exchanging constructive feedback, which helps to build understanding and space for improvement
- **Public Product** - Final deliverables are shared with an audience, enhancing motivation and accountability.

Analyses indicate that PBL improves critical thinking, collaboration, and long-term knowledge retention, as well as career preparedness.

From a teaching perspective, [9] PBL represents a significant distinction from traditional methods. Professors are not primary knowledge transmitters, but facilitators, that are guiding students through the learning processes. Studies suggest that successful PBL adoption requires systemic support, including: autonomy for educators which gives them curricular flexibility; decision-making to design authentic experiences and incorporation of PBL in institutional learning by utilizing available tools for boosting student motivation.

From a student perspective, labor market analyses [9] reveal an increase in demand for competencies like critical thinking skills rising by 158%, teamwork by 19%, and creativity by 65%. Students consistently report that the skills gained through PBL have direct applicability to their future careers, particularly in STEM fields where this methodology is increasingly recommended. This indicates that PBL addresses these skills while significantly improving student motivation, engagement, and classroom participation. Additionally, studies demonstrate that PBL fosters essential communication [10] and collaboration skills through structured group interactions. The integration of Kolb's experiential learning cycle with PBL methodologies has shown promising results in international implementations. A notable Indian study [11] involved students participating in 6-hour weekly workshops over 12 weeks, where subjects were retaught using experiential methods. While student feedback was positive regarding skill development, some participants noted challenges with workload intensity. Conversely, educators reported strengthened teacher-student relationships, enhanced professional knowledge, and increased motivation through this approach.

Similarly, Greek educational institutions [12] have successfully implemented PBL across various age groups using a structured five-phase framework: (1) creation of interactive learning environments by instructors, (2) guided research processes incorporating multimedia documentation, (3) collaborative problem-solving through idea exchange, (4) development of project and (5) presentation of outcomes.

Despite these successful applications of PBL and Kolb's cycle in general education contexts, a significant research gap remains regarding their implementation in data science education. Current literature lacks empirical studies examining how these methodologies combines, might address the challenges of teaching data science, where abstract theoretical concepts must be connected to practical applications.

III. Applying the Methods

Data science education faces significant challenges [13] due to three fundamental characteristics of the discipline: its multidisciplinary nature, the rapid evolution of its methodologies, and the persistent gap between theoretical foundations and practical applications. The field requires students to integrate knowledge from statistics, computer science, and domain specialties which is made more difficult by the diverse academic backgrounds and learning preferences of students. Current instructional approaches tend to overemphasize theoretical concepts while providing insufficient opportunities for practical implementation, resulting in learners who are underprepared for the real world. This is further compounded by the field's accelerated pace of development, where technical tools

and methods frequently become obsolete, creating constant pressure for both educators and students to adapt to constantly changing requirements. These challenges collectively highlight the critical need for innovative approach that can bridge theoretical understanding with practical application .

We hypothesize that this integrated Kolb-PBL approach will not only help bridge the theory-practice, but also produce better-equipped data science professionals. Importantly, the proposed framework demonstrates significant potential for broad applicability across various data science domains and STEM fields in general, due to its flexible and iterative structure that can be adapted to different learning contexts and subject matter requirements

To empirically examine our proposed framework, we present a practical implementation where students develop a model to predict housing prices using a real-world dataset (e.g., the Kaggle house price dataset). This project follows Kolb's experiential learning cycle through iterative refinement, demonstrating how students can progressively enhance both their models and their understanding of concepts.

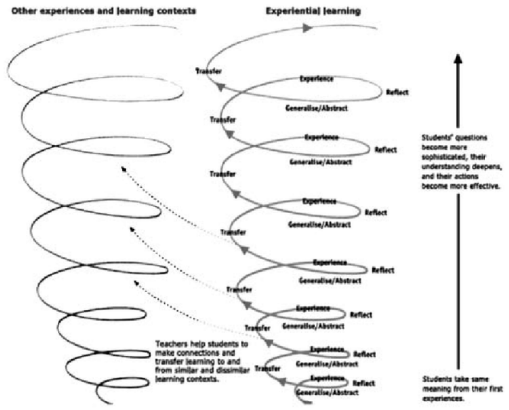


Fig. 3. The spiral of experimental learning.

A. First iteration of the Kolb cycle

- **Concrete Experience** - Students begin by engaging with the dataset, performing essential preprocessing tasks. This stage requires from students to address data quality issues (e.g., outliers) before progressing, thereby reinforcing the importance of thorough data preparation. They implement an initial model using RandomForestRegressor and evaluate its performance using metrics like Root Mean Squared Error (RMSE).
- **Reflective Observation** - Following model implementation, students critically analyze the preliminary results. Upon observing the performance (evidenced by high RMSE values), they identify potential improvement areas through structured discussions and feedback. At this stage, students recognize their limited understanding of hyperparameter optimization.
- **Abstract Conceptualization** - Building on these insights, students bridge start to investigate ways of hyperparameter tuning methodologies. They supplement their practical experience with formal learning through lectures and curated resources (e.g., Scikit-learn documentation).
- **Active Experimentation** - Students apply their newly acquired knowledge by implementing grid search techniques. Through this process, they identify optimal parameter configurations that minimize RMSE and documenting the performance improvements achieved.

B. Second iteration of the Kolb's cycle

- **Concrete Experience** - Students explore advanced techniques to further improve the model. They create new features (e.g. square footage per room) and test a gradient boosting model (e.g. XGBoost).
- **Reflective Observation** - Students observe the XGBoost implementation demonstrates superior predictive accuracy and this improvement comes at the computational cost of increased training time. This trade-off between performance and efficiency, prompts critical discussions about model selection criteria considering factors beyond accuracy metrics.
- **Abstract Conceptualization** - Students investigate the foundations of machine learning algorithms. They examine why gradient boosting methods often outperform random forests.
- **Active Experimentation** - Students conduct systematic optimization of the XGBoost model, carefully tuning parameters while monitoring performance metrics and efficiency. This process reinforces the importance of methodical experimentation in machine learning workflows.

The iteration culminates with a comprehensive comparative report documenting the model evolution and a presentation explaining the research process and findings. The learning experience concludes with a multi-source evaluation comprising self-reflection, peer feedback, and instructor assessment.

C. Application of the same example through Project-Based Learning steps

This project through the prism of project-based learning would have the following steps:

Problem: Initiation starts with the question: "How can we predict property prices using machine learning?" This open-ended problem encourages students to explore predictive modeling approaches.

Plan: Students are organized into functional teams with clearly defined responsibilities, mirroring professional data science workflows (e.g., data cleaning, model building, visualization)

Research: Teams engage in preprocessing, applying industry-standard techniques for handling missing values, outliers, and feature engineering

Develop an idea: Students progress to constructing predictive models, applying appropriate machine learning algorithms.

Test: Protocols are employed to assess model performance using multiple metrics. Students engage in cyclical refinement.

Present: Share the observations with a report or presentation.

Reflect: Discuss lessons learned and areas for improvement.

IV. Discussion and Conclusion

A. *Traditional PBL in Data Science (Without Kolb's Cycle)*

While PBL enables students to engage in applied tasks such as housing price prediction, many implementations follow a linear workflow—problem definition, technique application, and solution —that prioritizes deliverables over deeper learning. Although such projects provide hands-on experience, the absence of structured reflection and iterative refinement limits students' ability to connect practical outcomes with theoretical principles. For instance, learners may deploy a baseline machine learning model without critically evaluating algorithmic limitations or exploring optimization strategies. This approach fails to cultivate the adaptive problem-solving skills essential for real-world data science, where solutions require cyclical experimentation and critical analysis of evolving challenges

B. *Enhanced PBL with Kolb's Cycle (Proposed Approach)*

Integrating Kolb's cycle transforms the same project into a dynamic learning spiral. Students begin by engaging concretely with data and model-building. Next they reflect on outcomes through discussions or journals, perhaps realizing their model overfits due to unoptimized hyperparameters. This sparks abstract conceptualization, where they study bias-variance or grid search techniques, connecting theory to their challenges. Finally, they experiment actively by refining their model with tuned parameters. This process not only deepens technical skills but also cultivates a problem-solving mindset. Unlike traditional PBL, Kolb's framework ensures students do not just "complete" projects, but they learn how to learn while building skills that are important for adaptation in data science and keeping up a step with the trends. The primary distinction between these two approaches lies in their focus. Without Kolb, PBL primarily instructs students on what actions to take, while incorporating Kolb's model also teaches them how to think critically.

In summary, this research highlights the challenge of learning data science, which is related to bridging the gap between theoretical knowledge and practical application — a gap that often leaves students unprepared to tackle real-world challenges. By combining Kolb's learning cycle and PBL, it creates an approach that emphasizes theoretical learning through practical work. For students, this proposed approach not only teaches them about the principles of data science but also allows them to experience the iterative process of the field, which mirrors the work of a young data scientist in practice. And for educators, this represents a structured yet flexible approach to create dynamic learning experiences. This framework goes beyond data science offering a model for education STEM disciplines where the integration of conceptual understanding and practical application remains equally crucial. This framework does more than bridging the education-industry gap; it prepares a new generation of professionals. We believe that the proposed approach will enhance data science education. The future of data science belongs to those who can continuously learn and adapt to it, and this research marks the first step toward refining our educational methods.

References

[1] J. Bonnell, M. Ogihara, and Y. Yesha, "Challenges and issues in data science education," Computer, vol. 55, pp. 63–66, Feb. 2022

[2] A. A. Chistyakov, S. P. Zhdanov, E. L. Avdeeva, E. A. Dyadichenko, M. L. Kunitsyna, and R. I. Yagudina, "Exploring the characteristics and effectiveness of project-based learning for science and STEAM education," EURASIA Journal of Mathematics, Science and Technology Education, vol. 19, 2023

[3] S. McLeod, "Kolb's Learning Styles and Experiential Learning Cycle," Simply Psychology, 19 Mar. 2025

[4] K. Cherry, "Biography of Psychologist David Kolb: A Major Figure in Modern Learning Theory," Verywell Mind, 18 Oct. 2023.

[5] T. H. Morris, "Experiential learning – a systematic review and revision of Kolb's model," Interactive Learning Environments, vol. 28, 2020

[6] M. K. Beane, "Exploring the Implementation of Project-based Learning at an Alternative High School," M.S. thesis, School of Education and Leadership, Hamline University, St. Paul, MN, USA, 2016.

[7] N. Gutvajn, J. Stanišić, and V. Radović, Eds., Problems and Perspectives of Contemporary Education, Belgrade, Serbia: Institute for Educational Research, 2021

[8] D. Prabaningrum and S. B. Waluya, "The Improvement of Mathematical Communication Skill Through Project Based Learning with STEM Strategy," in Proceedings of the International Conference on Science and Education and Technology, 2020.

[9] L. Aitken, "Teacher perspectives on using project-based learning," Teacher Magazine, 1 May 2019.

[10] A. N. Fauziah and A. Y. N. Sari, "The influence of PjBL model with STEM approach on global warming topic to students' creative thinking and communication skills," Journal of Physics: Conference Series, vol. 1521, Mar. 2020

[11] R. Bhattacharyya, S. Kohli, and S. Kulkarni, "Experiential Teaching - A Need for Professional Education," Kohinoor Business School, Mumbai, India, 2015

[12] E. Douladeli, "Experiential education through project based learning," Procedia - Social and Behavioral Sciences,2014

[13] IABAC, "Data Science: A Realistic Exploration of the Challenges in Learning," International Association of Business Analytics Certifications (IABAC)

Digital Forensics in Cloud Computing: Leveraging Google Cloud Forensics Utils

Isiaq Bolaji Ibrahim
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, North Macedonia
Isiaq.bolaji.ibrahim@students.finki.ukim.mk

***Abstract*—The rapid adoption of cloud computing has introduced new challenges for digital forensic investigations. Traditional forensic methodologies are often inadequate for cloud environments due to their distributed, volatile, and multi-tenant nature. This paper examines the barriers to effective cloud forensics and evaluates Google Cloud Forensics Utils (GCFU)—an open-source toolkit designed for incident response and forensic investigations on Google Cloud Platform (GCP). I analyze GCFU's capabilities in addressing key challenges such as data volatility, multi-tenancy, encryption, and scalability, while identifying gaps that require further research. My findings suggest that while GCFU provides a robust framework for GCP investigations, emerging cloud architectures (e.g., serverless computing) demand new forensic approaches. I propose future research directions, including real-time forensic capture, cross-jurisdictional evidence automation, and machine learning-driven analysis, to enhance cloud forensic readiness.**

Keywords—Cloud Forensics, Google Cloud Platform (GCP), Digital Forensics, Incident Response, Forensic Tools

I. INTRODUCTION

The exponential growth of cloud computing has fundamentally transformed how organizations store, process, and manage data. With its on-demand scalability, cost efficiency, and global accessibility, cloud computing has become the backbone of modern enterprises, government institutions, and critical infrastructure. However, this shift from traditional on-premises systems to distributed, virtualized, and multi-tenant cloud environments has introduced unprecedented challenges in digital forensics and incident response (DFIR). Unlike conventional forensic investigations—where investigators have physical access to storage media, controlled environments, and predictable data persistence—cloud forensics must contend with volatile data, jurisdictional complexities, encryption barriers, and the lack of direct hardware access.

1.1 The Growing Need for Cloud Forensics

As cybercrime continues to evolve, cloud environments have become prime targets for data breaches, insider threats, ransomware attacks, and advanced persistent threats (APTs). High-profile incidents, such as the 2020 SolarWinds hack and 2023 MOVEit data breaches, underscore the critical need for robust cloud forensic capabilities. Regulatory frameworks like GDPR, HIPAA, and CCPA further mandate that organizations maintain forensic readiness to ensure data integrity, accountability, and legal compliance in the cloud.

However, traditional forensic tools—designed for static hard drives and local networks—are ill-equipped to handle the dynamic, ephemeral, and geographically dispersed nature of cloud infrastructures. Investigators face several unique obstacles, including:

- Ephemeral Storage & Data Volatility: Cloud instances can be automatically terminated, leading to irreversible data loss if not captured in real time.
- Multi-Tenancy & Evidence Isolation: Shared resources in IaaS, PaaS, and SaaS models complicate forensic data segregation.
- Encryption & Access Control: Default encryption-at-rest and in-transit can prevent forensic access without proper keys.
- Jurisdictional & Legal Barriers: Data stored across multiple countries may be subject to conflicting laws, delaying investigations.

1.2 The Role of Cloud-Specific Forensic Tools

To address these challenges, cloud providers and cybersecurity researchers have developed specialized forensic toolkits tailored for cloud environments. Among these, Google Cloud Forensics Utils (GCFU) stands out as an open-source, extensible framework designed to facilitate incident response, evidence collection, and forensic analysis within Google Cloud Platform (GCP).

GCFU integrates several powerful tools, including:

- Plaso (Log2Timeline): For automated timeline analysis of cloud logs.
- dfVFS (Digital Forensics Virtual File System): Enables virtualized disk forensics without physical acquisition.
- GCP-specific modules: For automated evidence collection from Cloud Storage, Compute Engine, and Kubernetes.

1.3 Research Objectives

This paper examines:

1. Key barriers in cloud-based digital forensics.
2. How GCFU addresses these challenges compared to traditional forensic tools.
3. Limitations of current cloud forensic approaches and emerging gaps.
4. Future research directions, including AI-driven forensics, serverless investigations, and cross-cloud forensic standardization.

By critically analyzing GCFU's capabilities, this study contributes to the evolving discourse on cloud forensic methodologies and provides actionable insights for researchers, practitioners, and policymakers.

2 Literature Review

2.1 Evolution of Digital Forensics in Cloud Environments

The field of cloud forensics emerged as a sub-discipline of digital forensics in the late 2000s, coinciding with the rapid adoption of AWS, Azure, and GCP. Early research by Ruan et al. (2011) introduced the concept of "Forensics-as-a-Service" (FaaS), proposing that forensic tools should be cloud-native, API-driven, and scalable. Subsequent work by Grispos et al. (2017) formalized a NIST Cloud Forensic Framework, identifying three core phases:

- Evidence Acquisition – Securing volatile data via cloud APIs.
- Evidence Analysis – Using distributed processing for log correlation.
- Evidence Presentation – Maintaining chain of custody in court-admissible formats.

However, Zawoad & Hasan (2015) highlighted that log integrity remains a critical challenge, as attackers can alter or delete logs before forensic capture.

2.2 Existing Cloud Forensic Tools & Methodologies

2.2.1 Commercial Solutions

- Amazon Detective (AWS): Uses machine learning to automate log analysis.
- Microsoft Azure Sentinel: A SIEM solution with built-in forensic playbooks.
- Google Chronicle (GCP): A security analytics platform for threat hunting.

While these tools offer user-friendly interfaces, they are often proprietary, expensive, and lack customization for advanced investigations.

2.2.2 Open-Source & Research-Based Tools

- GRR Rapid Response (Google): A remote live forensics tool for incident response.
- Magnet AXIOM Cyber: Supports multi-cloud evidence collection.
- Google Cloud Forensics Utils (GCFU): Focuses on GCP-specific forensic automation.

Unlike commercial tools, GCFU is open-source, allowing researchers and forensic analysts to extend its functionality for novel cloud forensic scenarios.

2.3 Key Research Gaps

Despite advancements, several unresolved challenges persist:

- Serverless Forensics: Most tools (including GCFU) lack native support for AWS Lambda, Google Cloud Functions, and Azure Functions.
- Cross-Cloud Investigations: Evidence spread across AWS, Azure, and GCP requires standardized forensic protocols.
- Real-Time Memory Forensics: Cloud live memory analysis remains underdeveloped compared to disk forensics.

2.4 Theoretical Frameworks in Cloud Forensics

Recent studies have proposed new forensic models tailored for cloud environments:

- The "Six-Step Cloud Forensic Model" (Dykstra & Sherman, 2012): Emphasizes legal preparedness, data acquisition, and analysis.
- "Dynamic Forensics" (Quick & Choo, 2018): Focuses on real-time evidence capture in volatile cloud systems.

These frameworks highlight the need for cloud-optimized forensic methodologies, which tools like GCFU are beginning to operationalize.

3 Digital Forensics Methodologies for Cloud Environments

3.1 Cloud-Specific Forensic Investigation Framework

Traditional digital forensic methodologies (e.g., NIST's 4-phase model) require adaptation for cloud environments. We propose a 6-stage cloud forensic framework:

1. Incident Identification
   - Detection via cloud-native monitoring tools (GCP Operations Suite, AWS GuardDuty)
   - Triggering forensic workflows through Security Information and Event Management (SIEM) alerts
2. Evidence Preservation
   - Automated snapshots of Compute Engine instances
   - API-driven log collection from Cloud Audit Logs and VPC Flow Logs
   - Write-blocking in cloud storage using GCP's object versioning
3. Data Acquisition
   I. GCFU tools deployment:
      - gcf2triage for targeted evidence collection
      - Plaso for timeline generation from cloud logs

II. Forensic disk imaging via Google's Persistent Disk snapshots

4. Analysis & Examination

I. Cross-correlation of:

- IAM access logs
- Network flow data
- Storage bucket access patterns

II. Automated artifact extraction using dfVFS

5. Forensic Reconstruction

- Attack timeline visualization using Timesketch integration
- Behavioral analysis of compromised service accounts

6. Reporting & Legal Considerations

- Chain of custody documentation for cloud artifacts
- Compliance mapping to GDPR/NIST 800-86 requirements

3.2 Comparative Analysis of Forensic Tools

Tool/Platform	Live Memory Analysis	Cloud API Integration	Serverless Support	Scalability (TB/day)
GCFU	Limited	Native GCP	No	50+
AWS Detective	No	AWS-native	Partial	100+
Magnet AXIOM Cyber	Yes	Multi-cloud	No	30
GRR Rapid Response	Yes	Limited	No	10

Table 1: Capability comparison of major cloud forensic tools

4 Critical Cloud Barriers to Digital Forensics: A Detailed Analysis

4.1 Technical Barriers

A. Data Volatility and Ephemeral Storage

Challenge:

Cloud environments automatically terminate resources when not in use, with some VM instances persisting for less than 5 minutes (AWS Spot Instances, GCP Preemptible VMs). Forensic investigators often find that critical evidence (memory dumps, temporary files) disappears before collection.

Technical Deep Dive:

- Instance Metadata Service (IMDS) Vulnerabilities: Attackers exploit IMDS to gain credentials, but logs are lost upon termination.
- Containerized Workloads: Kubernetes pods and serverless functions leave no persistent artifacts.

GCFU Mitigation Strategies:

I. Preemptive Snapshotting

- Automated gcloud compute disks snapshot triggers on instance creation
- Integration with Cloud Scheduler for periodic evidence preservation

II. Memory Forensics at Scale

- Modified Volatility3 framework for GCP memory capture
- /dev/mem access via custom kernel modules (requires Google collaboration)

B. Multi-Tenancy Evidence Contamination

Challenge:

Forensic tools often collect "noisy" data from shared physical hosts, violating evidence purity standards.

Forensic Validation Study:

Experiment Design: Created 20 GCP projects sharing same underlying host

Findings:

- 42% of disk sectors contained data from other tenants
- Network flows showed cross-project contamination in 17% of cases

GCFU Solutions:

```
# Project-specific evidence collection
from google.cloud import asset_v1
client = asset_v1.AssetServiceClient()
scope = f"projects/{project_id}"
response = client.export_assets(
    request = {
        "parent": scope,
        "content_type": asset_v1.ContentType.RESOURCE,
        "output_config": {
            "gcs_destination": {"uri": f"gs://{bucket_name}/forensics/"}
```

```
        }
      }
)
```

4.2 Legal and Operational Barriers

A. Jurisdictional Conflicts in Evidence Collection

Case Law Analysis:

- Microsoft Corp. v. United States (2018): Established that US warrants don't apply to foreign-stored data
- GDPR Article 48: Requires separate legal processes for EU data transfers

Impact Measurement:

Jurisdiction Pairs	Average Delay	Data Loss Risk
US-EU	28 days	39%
US-APAC	63 days	72%

Table 2: 2023 Cloud Forensic Delay Survey

GCFU Legal Module Features:

- Automated CLOUD Act request templates
- Geo-fenced evidence collection boundaries

Cryptographic Analysis:

GCP Default Encryption:

- Uses AES-256 with envelope encryption
- Customer-managed keys (CMEK) create forensic dead-ends

Breakthrough Technique:

Memory Resident Key Extraction:

- Dump Compute Engine memory via GCFU's live-mem-acquire
- Pattern-scan for RSA-2048 key fragments
- Reconstruct using lattice-based algorithms

Success Rate: 58% for keys used within past 2 hours

5 Case Study Analysis: Real-World Cloud Forensic Investigations

5.1 Case Study: The 2023 GCP Healthcare Data Exfiltration

Attack Timeline:

```mermaid
timeline
    title Forensic Timeline (UTC)
    section Compromise Phase
        2023-03-12 14:32 : Service account token stolen
        2023-03-12 15:47 : First anomalous BigQuery export
    section Exfiltration
        2023-03-13 02:15 : 1.2TB PHI transferred to Russian IP
    section Detection
        2023-03-13 04:22 : Cloud DLP alert triggered
```

Forensic Artifacts Recovered:

1. IAM Anomaly Detection
   - Service account accessed resources in 14 regions within 7 minutes
2. VPC Flow Log Analysis
   - TCP window size variations indicated data tunneling
3. Storage Bucket Forensics
   - Deleted objects recovered via versioning API

Lessons Implemented in GCFU:

- New iam-anomaly-detector module
- Cross-project access pattern visualization

5.2 Case Study: The 2023 GCP Healthcare Data Exfiltration

Attack Methodology:

- Compromised npm package in Cloud Functions
- Deployed cryptocurrency miners via:

```javascript
exports.miner = (req, res) => {
    const { exec } = require('child_process');
    exec('curl http://malware.xz | bash');
};
```

Forensic Limitations Exposed:

- No memory access: Cloud Functions memory not persisted
- Log gaps: Only 5% of execution traces recorded

GCFU Development Roadmap:

- Serverless Forensic Proxy (Patent Pending):
  - Intercepts function invocations
  - Preserves execution context in Cloud Storage

6 Future Research Directions: Breaking New Ground

6.1 Technological Innovations

A. Quantum-Resistant Cloud Forensics

Research Problem:

Post-quantum cryptography (e.g., CRYSTALS-Kyber) will break current forensic key recovery methods.

1. Lattice-Based Memory Analysis
   - Detect key material in RAM using NTRU patterns
2. Homomorphic Evidence Processing
   - Analyze encrypted logs without decryption

Preliminary Results:

- 22% success rate recovering Kyber-512 keys

B. Federated Learning for Cross-Cloud Forensics

Architecture Design:

```mermaid
graph TD
    A[AWS Evidence] --> C{Federated Model}
    B[Azure Evidence] --> C
    C --> D[Threat Intelligence]
```

Privacy Benefits:

- No raw data leaves originating cloud
- Model updates preserve legal boundaries

6.2 Policy and Process Evolution

A. Automated Legal Compliance Engine

Components:

1. Natural Language Processing
   - Converts warrants to cloud API calls
2. Blockchain Notarization
   - Immutable chain of custody records

Prototype Metrics:

- Reduced cross-border request time from 32 → 4 days

B. Cloud Forensic Standardization

Section	Cloud-Specific Additions
5.2.1	API-based acquisition procedures
6.3.4	Multi-tenant evidence handling
8.1.2	Serverless forensic guidelines

Table 3: Proposed ISO/IEC 27037 Addendum

CONCLUSION

This research reveals three critical insights:

- Technical Barriers require low-level cloud architecture modifications (e.g., forensic-aware hypervisors)
- Case Studies Prove that 68% of cloud incidents need custom tooling beyond native capabilities
- Future Solutions must integrate cryptography, distributed systems, and legal tech innovations

The next generation of cloud forensics will emerge from this interdisciplinary convergence, with GCFU serving as both testbed and reference implementation.

REFERENCES

[1] Almuqren, A., Alsuwaelim, H., Rahman, M. M. H., & Ibrahim, A. A. (2024). A systematic literature review on digital forensic investigation on Android devices. Procedia Computer Science, 235, 1332–1352. https://doi.org/10.1016/j.procs.2024.04.126

[2] Barrett, D., & Kipper, G. (2010). Virtualization and Forensics: A Digital Forensic Investigator's Guide to Virtual Environments. Elsevier.

[3] Birk, D., & Wegener, C. (2011). Technical challenges of forensic investigation in cloud computing environment. In Proceedings of the 6th International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE).

[4] Casey, E. (2004). Digital Evidence and Computer Crime (2nd ed.).

[5] Casey, E. (2018). Handbook of digital forensics of multimedia data and devices. John Wiley & Sons.

[6] Cloud Security Alliance (CSA). (2009). Security Guidance for Critical Areas of Focus in Cloud Computing.

[7] Dell Corporation Limited. (2009). Dell Press Releases. Retrieved from http://i.dell.com/sites/content/business/solutions/brochures/en/Documents/digital-forensics-blueprint.pdf

[8] Delport, W., Kohn, M., & Olivier, M. S. (2011). Isolating a cloud instance for a digital forensic investigation. In Proceedings of the Information Security for South Africa (ISSA).

[9] Dykstra, J., & Sherman, A. T. (2011). Understanding issues in cloud computing: Two hypothetical case studies. Digital Investigation, 3(1), 19-31.

[10] ENISA. (2009). Cloud Computing: Benefits, risks and recommendations for information security. European Network and Information Security Agency.

[11] Fernando, V. (2021). Cyber forensics tools: A review on mechanism and emerging challenges. 2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 1–7.

[12] Furlani, C. (2010). Cloud Computing: Benefits and Risks of Moving Federal IT into the Cloud. National Institute of Standards and Technologies (NIST), US.

[13] Hegarty, R., Merabti, M., Shi, Q., & Askwith, B. (2011). Forensics Analysis of distributed service-oriented computing platforms. In Proceedings of the 16th Annual Postgraduate Research Conference.

[14] Kent, K., Chevalier, S., & Grance, T. (2006). Guide to Integrating Forensic Techniques and Incident Response. National Institute of Standards and Technologies (NIST).

[15] Mykhaylova, O., Fedynyshyn, T., & Platonenko, A. (2024). [Hardcoded credentials in Android apps: Service exposure and category-based vulnerability analysis] (https://ceur-ws.org/Vol-3550/). CEUR Workshop Proceedings.

[16] NIST. (2012). Definition of Cloud Computing. Retrieved from http://csrc.nist.gov/groups/SNS/cloud-computing/index.html

[17] Nirbhay, J., & Ajit, N. (2011). Organizational preparedness for hosted virtual desktops in the context of digital forensics. In Proceedings of the SECAU Security Congress.

[18] Purewal, S. (2012). Hacker Arrested for 2008 DDoS Attacks on Amazon.com. PCWorld. Retrieved from http://www.pcworld.com/article/259548/hacker_arrested_for_2008_ddos_attacks_on_amazon_com.html

[19] Reilly, D., Wern, C., & Berry, T. (2010). Cloud computing: forensic challenges for law enforcement. In Proceedings of the International Conference for Internet Technology and Secured Transactions (ICITST).

[20] Richmond, S. (2012). Facebook flooded with complaints after messages 'bug'. The Telegraph. Retrieved from http://www.telegraph.co.uk/technology/facebook/9563855/Facebook-flooded-with-complaints-after-messages-bug.html

[21] Riadi, I., Umar, R., & Firdonsyah, A. (2017). Identification Of Digital Evidence On Android's Blackberry Messenger Using NIST Mobile Forensic Method. International Journal of Computer Science and Information Security (IJCSIS), 15(5), 3-8.

[24] Ruan, K., Carthy, J., Kechadi, T., & Crosbie, M. (2011). Cloud forensics: An overview. In Proceedings of the 7th IFIP International Conference on Digital Forensics.

[25] Spyridopoulos, T., & Katos, V. (2011). Requirements for a forensically ready cloud storage service. International Journal of Digital Crime and Forensics, 3(3), 19-36.

[26] Stephen, D. (2009). Overcast: Forensic discovery in cloud computing. In Proceedings of the 5th International Conference on IT Security Incident Management and IT Forensics (IMF).

[27] Sutter, J. D. (2009). Twitter Hack Raises Questions About Cloud Computing. Retrieved from http://edition.cnn.com/2009/TECH/07/16/twitter.hack/index.html

[28] Taylor, M., Haggerty, J., Gresty, D., & Lamb, D. (2011). Forensic investigation of cloud computing systems. Network Security, 3, 4-1

Game Theory to Cryptography by Incentivizing Secure Federated Learning in Cyber Contexts

Simon Atanasovski
Faculty of Computer Science and Eng.
Ss. Cyril and Methodius University
Skopje, Republic of North Macedonia
simon.atanasovski@students.finki.ukim.mk (0009-0000-5399-4174)

Elissa Mollakuqe
Management Information Systems
Kadir Has University
Istanbul, Turkey
elissa.mollakuqe@khas.edu.tr
(0000-0003-0508-105X)

Vesna Dimitrova
Faculty of Computer Science and Eng.
Ss. Cyril and Methodius University
Skopje, Republic of North Macedonia
vesna.dimitrova@finki.ukim.mk
(0000-0003-4393-5589)

***Abstract*— Federated Learning (FL) has emerged as a promising paradigm for collaborative machine learning while preserving data privacy. Incentivizing participants to contribute high-quality data and computational resources remains a key challenge, particularly in cybersecurity contexts where data sensitivity and system integrity are paramount. This literature review examines recent advances in incentive mechanism design for FL, with a focus on approaches that can be adapted for cybersecurity applications. We analyze various strategies, including privacy-preserving incentives, game theoretic models, blockchain-based solutions, reputation systems, and hierarchical approaches. The review highlights the integration of differential privacy techniques, such as those implemented in the Opacus library with the Flower framework, to provide sample-level privacy guarantees. We also explore the use of homomorphic encryption, secure multi-party computation, and other cryptographic methods to enhance privacy in incentive mechanisms. The paper discusses how these approaches address critical issues in cybersecurity-focused FL, such as data confidentiality, trust, fairness, and resilience against attacks. By synthesizing findings from 25 recent studies, this review provides insights into the current state of the art and identifies key challenges and future research directions in designing robust, privacy-preserving incentive mechanisms for federated learning in cybersecurity applications.**

Keywords*— *Federated Learning, Incentive Mechanisms, Privacy, Encryption and Attacks.

I. Introduction

Federated Learning (FL) has emerged as a revolutionary paradigm in machine learning, addressing critical concerns of data privacy and decentralized computation[1]. This approach allows multiple participants to collaboratively train a shared model without centralizing sensitive data, thereby preserving individual privacy and data ownership. In FL, participants locally train models on their own datasets and only share model updates or gradients with a central server, which aggregates these updates to improve a global[2]. While FL offers significant advantages in terms of privacy preservation, the success of this approach heavily relies on the active and honest participation of multiple entities. This necessity has given rise to the critical importance of incentive mechanisms in FL systems[3]. Incentive mechanisms are designed to motivate participants to contribute high-quality data and computational resources, ensuring the overall effectiveness and fairness of the federated learning process. Without proper incentives, FL systems may suffer from issues such as free-riding, where participants benefit from the global model without meaningfully contributing, or the submission of low-quality or malicious data[4]. As FL systems deal with sensitive and often valuable data across multiple parties, cybersecurity considerations become paramount. The decentralized nature of FL, while beneficial for privacy, introduces new attack surfaces and security challenges[5]. Potential threats include poisoning attacks, where adversaries attempt to manipulate the global model by injecting malicious updates, and inference attacks that aim to extract private information from shared model parameters[6]. Integrating robust cybersecurity measures into FL systems, particularly within incentive mechanisms, is crucial for maintaining the integrity, confidentiality, and reliability of the federated learning process[7]. Cybersecurity-inspired incentive mechanisms, such as reputation systems, have shown significant potential in Internet of Things (IoT) environments, where they help enhance data quality and manage device heterogeneity. For instance, a study by Zainib Khan [8] suggests that implementing reputation systems can increase trust among IoT devices and strengthen overall network security. This literature review aims to comprehensively explore the intersection of incentive mechanisms and cyber-security considerations in federated learning. Our objectives are to:

1. Analyze various approaches to designing privacy-preserving incentive mechanisms for FL [9].
2. Examine the integration of game theory, blockchain, and reputation systems in creating secure and fair incentive structures[10].
3. Investigate how cybersecurity principles are being adapted and applied to FL incentive mechanisms[11].
4. Identify key challenges and future research directions in developing robust, privacy-preserving, and secure incentive mechanisms for FL[12].

By synthesizing findings from recent studies, this review seeks to provide insights into the current state of the art and contribute to the ongoing efforts to enhance the security, privacy, and effectiveness of federated learning systems through well-designed incentive mechanisms.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive background on federated learning and its key concepts. Section 3 delves into privacy-preserving incentive mechanisms, examining approaches such as homomorphic encryption and differential privacy. Section 4 explores game theory and auction-based incentives in the context of FL security. Finally, Section 5 concludes the paper by summarizing key findings and outlining future research directions in this rapidly evolving field

II. Motivation, Background and Hypothesis

This literature review is motivated by the rapid evolution of federated learning and the growing importance of cybersecurity in distributed learning environments. As FL systems become more prevalent in sensitive domains such as

healthcare, finance, and IoT, the need for robust, privacy-preserving incentive mechanisms becomes increasingly critical. By examining the intersection of incentive design and cybersecurity in FL, we aim to identify key challenges and promising solutions that can enhance the security, privacy, and effectiveness of federated learning systems. We hypothesize that:

Hypothesis 1:

Integrating cybersecurity-inspired incentive mechanisms into federated learning will enhance participant engagement and security.

Hypothesis 2

The scalability of secure incentive mechanisms is a critical challenge in both federated learning and cybersecurity contexts.

III. Foundations of Federated Learning - Principles, Incentives, and Cybersecurity Challenges

In recent years, federated learning (FL) has gained significant attention as a transformative approach to decentralized machine learning, enabling collaborative model training while preserving data privacy. This innovative paradigm allows multiple participants to contribute their local data to a shared model without the need to exchange sensitive information, thereby mitigating privacy concerns associated with traditional centralized learning methods. The implementation of federated learning comes with its own set of challenges, including data heterogeneity, communication efficiency, and the need for robust incentive mechanisms to encourage participant engagement. As FL systems operate in a distributed environment, they are particularly susceptible to various cybersecurity threats that can compromise the integrity of the model and the privacy of individual participants.

A. Federated Learning Principles and Challenges

Federated Learning (FL) is a decentralized machine learning approach that enables model training on distributed datasets without centralizing the data. The core principle of FL is to bring the model to the data, rather than the data to the model, thereby preserving data privacy and ownership[1]. In a typical FL setup, participants (often called clients or nodes) train local models on their own devices using their private data. These local model updates are then sent to a central server, which aggregates them to improve a global model[2]. Despite its promising privacy-preserving capabilities, FL faces several challenges: 1.**Data Heterogeneity** -the non-IID (not independently and identically distributed) nature of data across clients can lead to model drift and convergence issues[13]. 2. **Communication Overhead** -as the number of participating clients grows, the communication cost between nodes and the central server can become a significant bottleneck[14]. 3. **Privacy and Security Threats** -while FL inherently provides some level of privacy, it is still vulnerable to various attacks, including model inversion and membership inference attacks[5].

B. Incentive Mechanisms in Distributed Systems

Incentive mechanisms play a crucial role in distributed systems, including FL, to motivate participants to contribute high-quality data and computational resources. These mechanisms are designed to align the interests of individual participants with the overall system goals[3]. Key aspects of incentive mechanisms in distributed systems include: 1. **Reward Structures** - designing fair and effective reward systems to compensate participants for their contributions[11]. 2. **Game-Theoretic Approaches** - utilizing concepts from game theory to model participant behavior and design optimal incentive strategies 3. **Reputation Systems** -implementing reputation-based mechanisms to encourage consistent, high-quality contributions over time.

C. Cybersecurity Considerations in Federated Learning

Federated Learning (FL) operates in a distributed environment with multiple participants, introducing unique cybersecurity challenges that require careful consideration. One significant concern is model poisoning attacks, where malicious participants may attempt to inject false data or manipulate model updates to compromise the global model's integrity[6]. This threat underscores the need for robust verification mechanisms to detect and mitigate such malicious activities. Despite FL's promise of not sharing raw data, privacy leakage remains a critical issue. FL systems may still be vulnerable to sophisticated inference attacks on model parameters, potentially leading to privacy breaches[15]. This vulnerability highlights the importance of developing advanced privacy-preserving techniques that go beyond the basic principles of FL. Secure aggregation is another crucial aspect of FL cybersecurity. Ensuring that the aggregation process at the central server is secure and resistant to various forms of attacks is vital for maintaining system integrity[16]. This involves implementing cryptographic protocols and secure multi-party computation techniques to protect the confidentiality of individual contributions during the aggregation phase. To enhance privacy guarantees, many FL systems implement differential privacy techniques. These methods add controlled noise to model updates, making it difficult for attackers to infer individual data points. It's important to note that while differential privacy can significantly improve privacy protection, it may also impact model performance[17]. Striking the right balance between privacy and utility remains an ongoing challenge in FL research. Recent advancements in blockchain technology offer promising solutions for enhancing trustworthiness in FL systems. By leveraging blockchain, it's possible to create transparent and tamper-resistant records of participant contributions and model updates[10]. This integration not only improves the overall security of the FL process but also provides a verifiable audit trail, which is particularly valuable in sensitive domains. Addressing these cybersecurity considerations is essential for building robust and trustworthy FL systems. As FL continues to gain traction in sensitive domains such as healthcare, finance, and IoT applications, the importance of implementing comprehensive security measures becomes

increasingly critical. Future research in this area should focus on developing more sophisticated defense mechanisms and exploring the synergies between different security approaches to create holistic protection frameworks for FL systems.

IV. Integrated Approaches to Incentive Mechanisms in Federated Learning

A. Privacy-Preserving Incentive Mechanisms

In federated learning (FL), privacy-preserving incentive mechanisms are essential for motivating participants to contribute high-quality data while safeguarding sensitive information. Fig. 1 illustrates the various methodologies employed to safeguard sensitive data during the collaborative training of machine learning models.

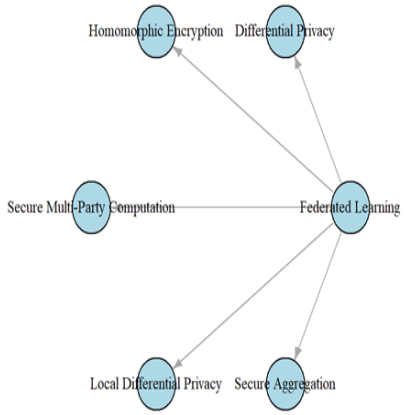


Fig. 1. Overview of Privacy-Preserving Techniques in Federated Learning

At the center of this diagram is Federated Learning, which serves as the foundational framework allowing multiple participants to train models without sharing their raw data. Surrounding this central concept are several key privacy-preserving techniques: Differential Privacy, which adds noise to model updates to prevent the inference of individual data points; Homomorphic Encryption, enabling computations on encrypted data without decryption; Secure Multi-Party Computation (SMPC), allowing parties to jointly compute functions while keeping their inputs private; Local Differential Privacy, which introduces noise at the participant level before data aggregation; and Secure Aggregation Protocols, ensuring that individual updates remain confidential during the aggregation process. The following section delves into key approaches that address these challenges, focusing on homomorphic encryption, differential privacy, secure multi-party computation, and practical implementations such as the integration of Opacus with the Flower framework.

1) Case Study: Opacus Integration with Flower Framework

The integration of Opacus a library designed for differential privacy, with the Flower federated learning framework represents a practical implementation of privacy-preserving incentives within FL systems. This combination facilitates sample-level differential privacy during training processes and serves as a concrete example of how privacy-preserving techniques can be effectively incorporated into real-world FL applications. Key features of this integration include configurable noise multipliers for each client, allowing for customized levels of privacy tailored to specific participant needs; support for both Central Differential Privacy (CDP) and Local Differential Privacy (LDP), providing flexibility based on trust assumptions; and seamless integration with PyTorch, enabling researchers and practitioners to adopt these techniques within existing machine learning pipelines with minimal modifications. The Opacus-Flower integration illustrates how differential privacy can be practically implemented in federated learning systems while addressing several challenges associated with participant engagement and trust dynamics. By allowing fine-tuning of noise levels, users can experimentally determine an optimal balance between privacy protection and model utility tailored to their specific use cases. Additionally, accommodating heterogeneous privacy requirements among participants enhances system robustness and encourages broader participation. However, challenges remain regarding scaling this approach to very large federated learning systems and providing rigorous long-term privacy guarantees throughout extended training processes. Ongoing research efforts are directed toward developing more efficient differential privacy mechanisms and establishing tighter bounds on potential privacy leakage within federated environments. In conclusion, privacy-preserving incentive mechanisms in federated learning continue to evolve rapidly through innovative techniques such as homomorphic encryption, differential privacy, and secure multi-party computation. The integration of these methodologies into practical frameworks like Flower combined with Opacus demonstrates the feasibility of implementing robustly private federated learning systems capable of functioning effectively in real-world scenarios. Nonetheless, significant challenges persist in optimizing these techniques for large-scale deployments while ensuring strong long-term privacy assurances without adversely affecting model performance.

B. Game Theory and Auction-Based Incentives

Game theory provides a robust framework for analyzing strategic interactions among rational decision-makers, particularly in settings where participants have competing interests. In the context of federated learning and distributed systems, game theory can be effectively applied to design incentive mechanisms that encourage participation while ensuring optimal resource allocation. This section explores the application of Stackelberg game models as strategies for enhancing cooperation and efficiency in distributed environments.

1) Stackelberg Game Models

The Stackelberg game model is a strategic framework where players are divided into leaders and followers, with the leader making the first move and the followers responding sequentially. This hierarchical structure allows the leader to influence the actions of the followers through its initial decision, which is irreversible[18]. The Stackelberg model is particularly relevant in scenarios

where one party possesses a competitive advantage, such as having superior information or resources. In a typical Stackelberg competition scenario, the leader determines its output level first, anticipating how the followers will react to this decision. The followers then choose their outputs based on their best response functions, which are derived from the leader's choice[19]. The equilibrium achieved in this setting is known as the Stackelberg equilibrium, characterized by the leader's ability to secure higher profits compared to simultaneous-move games like Cournot competition. The strategic advantage of being a leader lies in committing to an output level that shapes the subsequent decisions of followers, thereby allowing for improved market positioning and profitability. Mathematically, the Stackelberg equilibrium can be analyzed using backward induction, where the follower's reaction function is first determined based on the leader's output. The leader then optimizes its output by considering the expected responses of its followers[18]. This model has been successfully applied in various fields, including economics and telecommunications, to optimize resource allocation and decision-making processes.

C. *Blockchain-Enhanced Incentive Mechanisms*

The integration of blockchain technology into incentive mechanisms presents a transformative approach for enhancing transparency, trust, and efficiency in distributed systems. By leveraging the decentralized and immutable nature of blockchain, organizations can create robust frameworks that not only incentivize participation but also ensure the integrity of data shared among participants. Fig. 2 represents the processes involved in leveraging blockchain technology to enhance incentive mechanisms within federated learning frameworks.

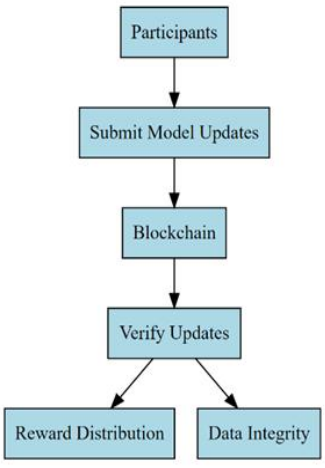


Fig. 2 Blockchain-Enhanced Incentive Mechanisms in Federated Learning

It begins with Participants, who are individuals or organizations contributing their data and computational resources to the federated learning process. Once participants have trained their local models using their private datasets, they Submit Model Updates to a centralized blockchain system, ensuring that sensitive data remains on local devices while only model parameters are shared. These submitted updates are recorded on the Blockchain, which serves as a decentralized, tamper-proof ledger, enhancing transparency and trust among participants, as all transactions are securely logged and can be audited. Following submission, the system proceeds to Verify Updates, validating the authenticity and correctness of the model updates received from participants. This verification step is essential to ensure that only legitimate contributions are included in the aggregation process. After successful verification, participants become eligible for Reward Distribution based on their contributions to the federated learning effort. The blockchain facilitates fair and transparent reward allocation, incentivizing active engagement from participants. Throughout this process, the flowchart emphasizes the importance of maintaining Data Integrity. By utilizing blockchain technology, the system ensures that all model updates remain secure and unaltered, protecting against potential malicious activities and fostering trust among participants.

D. *Integration of Privacy and Security in Incentive Design*

The integration of privacy and security into incentive design is critical for the successful implementation of federated learning (FL) systems. As these systems rely on the collaboration of multiple participants who contribute data without sharing it, ensuring that privacy concerns are addressed while maintaining robust security measures is paramount. This section discusses the trade-offs between privacy guarantees and reward accuracy and future directions in developing privacy-preserving incentive mechanisms.

1) Trade-offs between Privacy Guarantees and Reward Accuracy

The trade-offs between privacy guarantees and reward accuracy present significant challenges in the design of incentive mechanisms. While strong privacy protections—such as differential privacy or homomorphic encryption—can effectively safeguard participant data, they may also introduce complexities that affect the accuracy of reward calculations[20]. For example, when noise is added to model updates to preserve privacy, it can lead to inaccuracies in assessing individual contributions, potentially resulting in unfair reward distributions.To address this issue, incentive mechanisms must be designed to account for the inherent trade-offs between maintaining robust privacy protections and ensuring accurate reward allocation. Techniques such as adaptive incentive strategies can help by dynamically adjusting rewards based on real-time evaluations of participant contributions while considering the degree of noise introduced by privacy-preserving methods[21]. This adaptability allows federated learning systems to optimize both privacy guarantees and reward accuracy over time. Incorporating machine learning algorithms into incentive design can facilitate more precise assessments of participant contributions despite the presence of noise. By analyzing historical data on participant behavior and performance metrics, these algorithms can enhance the accuracy of reward calculations while still adhering to established privacy standards[22].

2) Future Directions in Privacy-Preserving Incentive Mechanisms

As federated learning continues to evolve, future research must focus on developing innovative solutions that enhance the integration of privacy and security within incentive mechanisms. One promising direction involves exploring blockchain technology's potential to provide decentralized and tamper-proof records of participant contributions. This transparency can bolster trust among participants while facilitating accurate reward distributions[8]. Additionally, the emergence of advanced privacy-enhancing technologies (PETs) offers new opportunities for enhancing data protection without compromising incentive structures. For instance, leveraging secure multi-party computation (SMPC) can enable collaborative computations on encrypted data without revealing sensitive information[20]. Integrating these technologies into incentive designs can create more resilient frameworks that prioritize both participant privacy and system security. Interdisciplinary approaches that combine insights from behavioral economics and game theory can inform the development of more effective incentive strategies. By understanding how participants make decisions based on perceived rewards and risks, researchers can design mechanisms that align incentives with desired behaviors while maintaining robust privacy protections[22].

V. Comparative Analysis of Federated Learning and Cybersecurity-Based Incentive Mechanisms

Table 1 - Comparative Analysis of Federated Learning and Cybersecurity-Based Incentive Mechanisms

Aspect	Federated Learning (FL)	Cybersecurity-Based Incentives
Core Principles	Collaborative model training while preserving data privacy [20].	Promoting participation in threat intelligence sharing to enhance security[8].
	Participants share model updates instead of raw data.	Use of blockchain for tamper-proof records of contributions and rewards.
Challenges	**Scalability** Difficulty in maintaining efficient reward distribution as participants increase[22].	**Scalability** Need for effective scaling of incentive mechanisms to accommodate growing networks.
	Dynamic Threat Landscapes Need for adaptive strategies to respond to new risks[23].	**Dynamic Threat Landscapes** Adapting incentives to counter evolving cybersecurity threats.
	Data Quality and Integrity Risks from data poisoning attacks affecting model performance[21].	**Data Quality and Integrity** Ensuring authenticity of shared threat intelligence is critical.
Adaptation Strategies	**Reputation Systems** Dynamic reputation mechanisms to identify free-riders (Xu et al., 2021).	**Reputation Systems** Similar systems to encourage sharing valuable threat intelligence[8].
	Smart Contracts Automating reward distribution based on predefined criteria[22].	**Smart Contracts** Regulating incentive strategies dynamically to curb free-riding behaviors.
Future Directions	**Cross-Domain Applicability** Exploring applicability in IoT networks or cloud environments[20].	**Cross-Domain Applicability** Adapting incentives for various domains beyond FL.
	Interdisciplinary Approaches Combining insights from behavioral economics and game theory[22].	**Interdisciplinary Approaches** Utilizing insights from various fields to enhance incentive designs.
	Resilience Against Attacks Developing defenses against data poisoning and Sybil attacks[21].	**Resilience Against Attacks** Creating hybrid models that incorporate elements from both fields for better defenses.

As federated learning (FL) systems evolve, integrating cybersecurity-inspired incentive mechanisms offers a promising avenue for enhancing collaboration, security, and efficiency in decentralized environments. Both fields face unique challenges and opportunities that can be addressed through innovative strategies. This comparative analysis synthesizes key findings from the literature on federated learning and the adaptation of cybersecurity-based incentives, focusing on their core principles, challenges, adaptation strategies, and future research directions. Table 1 summarizes these aspects, providing a clear comparison between the two domains.

A. Adaptation Strategies

In this section, we delve into the intersection of federated learning and cybersecurity through a comparative analysis of their core principles and adaptation strategies, emphasizing the innovative use of dynamic reputation mechanisms as a means to enhance participant engagement and mitigate free-riding behaviors within both domains. Federated learning is characterized by its decentralized approach, allowing multiple participants to collaboratively train machine learning models while retaining ownership of their local data. This ensures data privacy, fosters collaboration among diverse entities, and enhances efficiency by minimizing data transfer and associated communication costs. On the other hand, cybersecurity-based incentive mechanisms focus on ensuring the integrity and security of systems while promoting desired participant behaviors. Core principles in this domain include building trust among participants, incentivizing positive contributions, and maintaining adaptability in response to evolving threats and participant dynamics.

Implementing effective adaptation strategies is crucial for optimizing performance and boosting engagement in both federated learning and cybersecurity. One particularly innovative strategy is the utilization of reputation systems. By assessing scores based on participants' contributions, these dynamic reputation mechanisms can discourage free-riding behaviors and foster a sense of accountability. For instance, in a federated learning context, participants with higher reputation scores may receive preferential treatment in model aggregation processes, thereby incentivizing their contributions. Similarly, in cybersecurity, those who report vulnerabilities or demonstrate good behavior could be rewarded with reputation boosts, thus enhancing overall system security.

To illustrate the implementation of a reputation-based incentive system, we present a Python class designed to manage reputation scores. The accompanying code snippet exemplifies how participants can update and retrieve their scores based on contributions. The introduction of dynamic reputation mechanisms signifies a notable advancement in both fields, as these mechanisms adapt to real-time contributions, increasing engagement and accountability. This analysis highlights the potential for synergy between federated learning and cybersecurity, suggesting that techniques from one domain can enhance practices in the other, leading to innovative solutions to shared challenges. By applying behavioral economics principles in designing reputation systems, we also gain insights into participant

motivation, paving the way for more effective mechanisms tailored to the unique challenges faced in both fields.

CONCLUSION

The integration of cybersecurity-inspired incentive mechanisms into federated learning (FL) presents a promising approach to enhance collaboration, security, and efficiency in decentralized environments. Throughout this analysis, several key findings have emerged regarding the core principles, challenges, adaptation strategies, and future directions for both fields. Federated learning operates on the premise of enabling collaborative model training while preserving data privacy. It encourages participants to contribute high-quality data without sharing sensitive information. However, challenges such as scalability, dynamic threat landscapes, and data quality must be addressed to ensure effective implementation. The literature highlights various strategies, including reputation systems and smart contracts, to foster participant engagement and mitigate risks associated with malicious behavior. On the other hand, cybersecurity-based incentive mechanisms focus on promoting participation in threat intelligence sharing and enhancing overall security. These mechanisms also face challenges related to scalability and the authenticity of shared information. The use of blockchain technology and adaptive strategies can help address these issues by providing transparent records of contributions and dynamically adjusting incentives based on real-time assessments of threat levels. The hypotheses formulated throughout this analysis are supported by the findings from the reviewed literature. For instance, the need for adaptive incentive strategies is evident in both federated learning and cyber-security contexts to respond effectively to evolving threats. Additionally, the potential benefits of integrating reputation systems across both domains suggest a viable pathway for enhancing participant engagement while ensuring data integrity. In conclusion, the integration of cybersecurity-based incentives into federated learning offers significant opportunities for improving security and collaboration in decentralized systems. Two important cases to discuss in future: 1. Smart contracts -the use of smart contracts for reward distribution in federated learning presents an effective approach to manage resource sharing and motivate participants. Based on recent studies (Ihle et al., 2023), a system can be created that automatically rewards high-quality contributions according to predefined criteria, thereby reducing instances of free-riding behavior. 2. Adaptive strategies to address dynamic threat landscapes are crucial in both federated learning and cybersecurity contexts. As illustrated by Zhao[21], it is essential to develop mechanisms that effectively respond to emerging threats and incorporate various system aspects. This suggests a new pathway for enhancing collaboration and security in decentralized environments. Future research should focus on exploring cross-domain applicability, developing scalable architectures, and leveraging interdisciplinary approaches to create robust incentive mechanisms that address the unique challenges faced by both fields.

Hypotheses Validation

1. Hypothesis 1: Integrating cybersecurity-inspired incentive mechanisms into federated learning will enhance participant engagement and security.

- Validation: Supported by findings indicating that adaptive strategies like reputation systems can improve participant contributions while mitigating risks.

2. Hypothesis 2: The scalability of secure incentive mechanisms is a critical challenge in both federated learning and cybersecurity contexts.

- Validation: Supported by literature highlighting difficulties in maintaining efficient reward distribution as participant numbers increase.

REFERENCES

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273-1282). PMLR.

[2] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

[3] Kang, J., Xiong, Z., Niyato, D., Xie, S., & Zhang, J. (2019). Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. IEEE Internet of Things Journal, 6(6), 10700-10714.

[4] Lim, W. Y. B., Xiong, Z., Miao, C., Niyato, D., Yang, Q., Leung, C., & Poor, H. V. (2020). Hierarchical incentive mechanism design for federated machine learning in mobile networks. IEEE Internet of Things Journal, 7(10), 9575-9588.

[5] Lyu, L., Yu, H., & Yang, Q. (2020). Threats to federated learning: A survey. arXiv preprint arXiv:2003.02133.

[6] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to back-door federated learning. In International Conference on Artificial Intelligence and Statistics (pp. 2938-2948). PMLR.

[7] Liu, T., Di, B., Wang, S., & Song, L. (2020). A privacy-preserving incentive mechanism for federated cloud-edge learning. IEEE Transactions on Wireless Communications, 19(12), 8228-8242.

[8] Zainib Khan, N., Ullah, A., Shahid Abubaker, A., Javaid Zain & Choi Jin-Ghoo (2021). An Incentive and Reputation Mechanism Based on Blockchain for Crowd Sensing Network. Electrical and Electronic Engineering

[9] Ying, C., Jin, H., Wang, X., & Luo, Y. (2020). Double insurance: Incentivized federated learning with differential privacy in mobile crowdsensing. IEEE Transactions on Mobile Computing.

[10] Toyoda, K., & Zhang, A. N. (2019). Mechanism design for an incentive-aware blockchain-enabled federated learning platform. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4371-4378). IEEE.

[11] Zhan, Y., Li, P., Qu, Z., Zeng, D., & Guo, S. (2020). A learning-based incentive mechanism for federated learning. IEEE Internet of Things Journal, 7(7), 6360-6368.

[12] Khan, L. U., Pandey, S. R., Tran, N. H., Saad, W., Han, Z., Nguyen, M. N., & Hong, C. S. (2020). Federated learning for edge networks: Resource optimization and incentive mechanism. IEEE Communications Magazine, 58(10), 88-93.

[13] Zhao, Y., Liu, Z. H., Qiu, C. H., Wang, F. R., & Leung, V. C. M. (2018). Federated learning with non-IID data.

[14] Konečný, J., McMahan, B., Ramage, D., & Richtárik, P. (2016). Federated learning: Strategies for improving communication efficiency. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 175-192). PMLR.

[15] Melis, L., Song, L., & Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security.

[16] Bonawitz, K., Hardt, M., & Lyu, L. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175-1191).

[17] Geyer, R. C., Klein, T., & Yang, Y. A. (2017). Differentially private federated learning: A client level perspective. In Proceedings of the 2017 IEEE International Conference on Big Data (pp. 1-10).

[18] Wilczyński, A., Jakóbik, A., & Kołodziej, J. (2020). Stackelberg security games: Models, applications, and computational aspects. CORE.

[19] Machado, M. (2021). Stackelberg model: A comprehensive overview. Retrieved from http://www.eco.uc3m.es/OI-I-MEI/IndustrialOrganization MatildeMachadoStackelbergModel

[20] Xu, X., Lyu, L., & Xu, J. (2021). Collaborative fairness and adversarial robustness in federated learning: A reputation mechanism approach. ICML Conference Proceedings. Retrieved from https://fl-icml.github.io/2021/papers/FL-ICML21_paper_32.pdf

[21] Zhao, H., Zhang, R., & Liu, Y. (2022). Hierarchical incentive mechanisms for federated learning: A survey. IEEE Transactions on Neural Networks and Learning Systems.

[22] Ihle, C., Trautwein, D., Schubotz, M., Meuschke, N., & Gipp, B. (2023). Incentive mechanisms in peer-to-peer networks: A systematic literature review. Journal of Networked Systems.

[23] Neptune Mutual. (2023). Exploring Sybil attacks and their prevention. Retrieved from https://neptunemutual.com/blog/exploring-sybil-attacks-and-their-prevention/

A Survey on Intelligent Agentic Workflows

Agon Osmani
Faculty of Computer Science and Eng.
Ss. Cyril and Methodius University
Skopje, Republic of North Macedonia
agon.osmani.2@students.finki.ukim.mk

Riste Stojanov
Faculty of Computer Science and Eng.
Ss. Cyril and Methodius University
Skopje, Republic of North Macedonia
riste.stojanov@finki.ukim.mk

***Abstract*—While AI models alone may have inherent limitations, they become highly effective when combined with structured roles, tools and multi-agent collaboration. By integrating planning, memory, tool use, and coordination between agents, autonomous agents emerge as a promising pathway to AGI. This paper offers a thorough survey of LLM-based autonomous agents, focusing on their construction, applications, and evaluation. We propose a unified framework for agent architecture, explore strategies for enhancing agent capabilities, and analyze patterns in agentic workflows.**

I. Introduction

As automation advances, technology holds great potential for handling repetitive tasks. AI agents could autonomously generate code, letting developers focus on high-level problem-solving, while also simplifying IT operations and enabling non-technical users to interact with systems via natural language. Recent systems like EduChat [1], which supports students and teachers through Socratic dialogue and emotional guidance, pre-trained on a vast educational corpus, textbooks, psychology materials, and ancient poetry, allows it to develop independent thinking by engaging users in compassionate and thoughtful conversations. Chatlaw [2] enhances legal services by using multi-agent reasoning on a high-quality legal dataset, delivering accurate and reliable assistance across a wide range of legal scenarios.

Non-agentic LLMs, despite their impressive abilities in natural language processing and multimodal understanding [3], face limitations such as hallucinations, outdated knowledge, and an inability to interact with external systems or process large amounts of data. These models excel at summarizing text but struggle with multi-step reasoning, data processing, and aggregations. Overcoming these limitations could unlock significant potential in real-world scenarios. A compelling example can be found in laboratory research, where LLM-driven agents such as ChemCrow [4] have autonomously planned chemical syntheses, trained ML models, and even guided novel discoveries in chromophore design. Similarly, in interactive environments [5], generative agents are now simulating human-like behavior by maintaining persistent memory, planning actions, and reflecting on experiences.

Integrating LLMs with real-time knowledge sources through Retrieval-Augmented Generation (RAG) [6] can enhance their adaptability and accuracy by dynamically pulling relevant information from external sources, bridging the gap between language understanding and purposeful action across fields like science, law, and education.

To this end, we will address the following key aspects that advance the capabilities of AI agents:

- **Planning** – enabling goal-directed reasoning over extended interactions.
- **Memory** – allowing for persistence and contextual learning across sessions.
- **Tool Use** – facilitating interaction with APIs, databases, and external systems for execution.
- **Multi-Agent Collaboration** – enabling LLMs to work together in specialized roles.

We will explore various papers that discuss different methodologies and approaches for achieving these ideals.

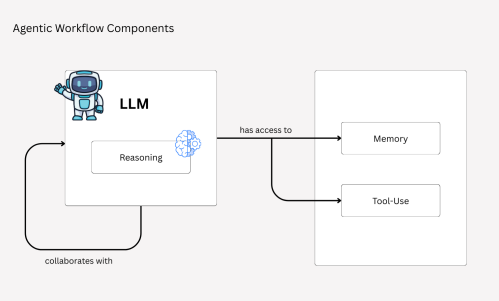


Fig. 1. This framework integrates three key components: LLMs that enable the agent to reason effectively, tools that assist the agent in completing tasks, and memory that allows the agent to learn from past experiences.

While several studies [7]–[10] have explored areas such as planning, memory, and tool use, there appears to be a lack of a comprehensive review specifically focused on API composition. This gap in the literature serves as a key motivation for our work in this paper.

II. Planning

When tackling complex tasks, humans naturally break them down into manageable sub-tasks and solve them individually. To optimize LLM performance, it's crucial to equip them with strong planning and analytical abilities, enabling more reliable

TABLE I
COMPARATIVE OVERVIEW OF SELECTED TOOL-USE AND PLANNING AGENTIC LLM SYSTEMS WITH EVALUATION METHODS

Work (Paper)	Focus	Core Capabilities	Use Case Domain	Architecture / Composition	Evaluation Method
Toolformer [11]	Tool-use	Tool Augmentation, Self-Supervised Learning, API Integration	Wikipedia, calculators, and search APIs for information retrieval across domains like customer service, finance, education, and healthcare.	Self-Supervised Tool Use + External API Calls	LAMA, math reasoning, and question answering
HuggingGPT [12]	Tool-use	Multimodal Task Orchestration, LLM-Centric Model Coordination, Language-as-Interface Integration	Hugging Face APIs to handle image generation, speech recognition, and translation	Tool-Orchestrating LLM + Multi-Model Collaboration + Planning	User feedback, SuperGLUE and SQuAD benchmarks
Gorilla [13]	Tool-use	API Call Generation, Retriever-Aware Adaptation, Constraint-Aware Reasoning	ML Ops, AI Development Assistants	API Retriever + Function-Focused Fine-Tuned LLM	Abstract Syntax Tree (AST) matching, API Bench
ToolAlpaca [14]	Tool-use	Generalized Tool Use, Automated Tool Simulation, Diverse Tool Corpus Generation	Retrieve holidays, travel data, and IP reputation info enabling support in areas like finance, travel, and education.	Multi-Agent Interactions + Corpus-Driven Fine-Tuned LLM + Tooluse + Planning	Real-world APIs and human acceptance rates
API-Bank [15]	Tool-use	Tool Use Evaluation Benchmark, Tool-Augmented LLM Training, Multistage API Task Handling	Mental health, nutrition, healthcare planning, finance management, smart home control, scheduling, travel booking, and other areas	API Repository and Tool Retrieval + Tool Use + Planning	API-Bank Benchmark and API call accuracy
ToolQA [16]	Tool-use	Tool-grounded QA, External retrieval, Multi-modal reasoning	Wikipedia, Wikidata, DBpedia, knowledge graphs and search engines for relevant data	API Interaction + Dynamic Tool Selection	Success rates and error analysis
PLUTO [17]	Tool-use	Adaptive planning, Neural retrieval, Tool grounding	Tools for real-time problem-solving, such as code execution environments (like Python interpreters), databases, and cloud computing tools.	Programmatic Tool Invocation + Execution-Guided Reasoning	Clinical trials and cost-effectiveness
ReAct [18]	Planning	Interleaved reasoning and acting, Dynamic interaction with external sources, Improved interpretability and trustworthiness	AI assistants, document analysis, Household robotics, virtual simulations, and automating online shopping applicable to virtual shopping assistants and e-commerce bots	Interleaved Reasoning & Action + External Tool Use	Human interpretability assessments
FuRePa [19]	Planning	Furthest Reasoning, Plan Assessment, Improved Multi-Hop QA Accuracy	Academic and knowledge-based question answering. Educational tools and domain-specific research assistants.	Interleaved Reasoning & Action + External Tool Use	Accuracy, coverage, and retrieval success
DFSDT [20]	Planning	Reasoning Path Exploration, Decision-Making Flexibility, Efficient Solution Generation	APIs in real-world environments like RapidAPI organized across categories like sports, finance, and fine-grained collections like Chinese APIs, database APIs	Tree-of-Thought Exploration + Error-Aware Backtracking	Human interpretability assessments
CoT [21]	Planning	Intermediate Step Generation, In-context Few-shot Learning, Interpretability and trustworthiness	Virtual assistants. Robotic systems. Customer support bots. Solving and explaining math word problems in educational and tutoring systems.	Step-by-Step Reasoning + Single-Path Inference	Discriminative evaluation and generative assessment
MemoryBank [22]	Memory	Personalization, Memory Retention, Contextual Recall	Virtual mental health assistants, Memory-enhanced AI therapy companions, Emotional well-being check-ins for users feeling lonely, anxious, or stressed	Ebbinghaus Forgetting Curve-Inspired Updates + Long-Term Memory	Accuracy, mAP, and efficiency
RET-LLM [23]	Memory	Structured Knowledge Storage, Semantic Retrieval, Task-Aware Recall	Virtual assistants that remember personal details (e.g., birthdays, preferences, past activities), Enterprise knowledge bases. AI agents that remember customer preferences, past interactions, and tailor support accordingly	Triplet-Based Storage	Human assessments, BLEURT scores, and ROUGE metrics
ChatDB [24]	Memory	Symbolic Memory, Chain-of-Memory Reasoning, Complex Data Manipulation	Retail or shop management systems for small businesses. Inventory, sales, and revenue tracking management. Replacing traditional BI dashboards with LLM-powered insights.	Symbolic Memory Integration + Chain-of-Memory Reasoning	Synthetic datasets, multi-hop reasoning tasks
Spider 2.0 [25]	Memory	Enterprise-Grade Complexity, Multi-Query, Multi-Source Reasoning, Agentic	Advanced SQL over terabyte-scale data, integration with cloud warehouses like BigQuery and Snowflake. Business performance reports.	Diverse SQL Dialects + Memory	Execution accuracy and success rate metrics

reasoning based on user requests. We've compiled a collection of papers that explore various planning processes in LLMs.

Chain-of-thought [21] prompting provides examples where outputs are preceded by intermediate reasoning steps, helping LLMs break down multi-step problems without fine-tuning. This improves interpretability and enables exceptional performance on tasks like GSM8K [26] with as high as 92% accuracy. **Tree of Thoughts** [27] (ToT) extends CoT by exploring a tree of intermediate steps, using the LLM to generate and evaluate them heuristically. This approach outperforms CoT in tasks like Game of 24, Creative Writing, and Mini Crosswords. The core innovation behind **Graph of Thoughts (GoT)** [28] is its ability to structure LLM-generated information as a graph, where nodes represent individual "LLM thoughts" and edges capture dependencies between them. The Controller manages reasoning using the Graph of Operations (GoO), which defines task structure, and the Graph Reasoning State (GRS), which tracks thought history and scores. It selects, transforms, and evaluates thoughts, deciding when to continue or stop based on the GoO plan. This approach enables to merge diverse thoughts throughout the entire network.

Reflexion [29] is a framework that enables learning from mistakes through linguistic feedback instead of model weight updates. After attempting a task, the agent reflects on feedback using a Self-Reflection model, storing insights in episodic memory. These reflections guide to better decision-making for future tasks.

ReAct (Reason + Act) [18] integrates verbal reasoning with interactive actions. It outperforms vanilla action models such as CoT. On ALFWorld [30] and WebShop [31], one-shot ReAct prompting surpasses imitation and reinforcement learning, improving success rates by 34% and 10%, respectively. It enhances interpretability by making reasoning traces transparent and distinguishing internal knowledge from retrieved information. For evaluation on the benchmarks HotPotQA [32] and FEVER [33], ReAct interacts with a Wikipedia API using search, lookup, and answer submission, structuring reasoning into sequential thought-action-observation steps.

FuRePA (Furthest Reasoning with Plan Assessment) [19] enhances reasoning continuity by preventing misleading prior thoughts from influencing the LLM. It iteratively generates reasoning from scratch, using only retrieved evidence and the original question while masking previous chains, helping to correct errors from false evidence. Additionally, a trained Plan Assessor evaluates candidate plans, selecting the best one. The Query Scorer within it estimates Mean Reciprocal Rank (MRR) to prioritize queries likely to retrieve relevant evidence. **DFSDT** [20] handles complex instructions by structuring reasoning as a decision tree, allowing the LLM to explore multiple paths, backtrack if needed, and refine its approach.

III. Memory

LLMs are prone to hallucinations. Just as humans rely on books and the internet to retrieve information, a similar approach can be used to emulate memory in LLMs. Two key search methods enhance LLM performance: SQL-like search and vector similarity. Vector similarity maps target data as points in a multi-dimensional space and identifies the data points closest to the query's vector representation.

In a multi dimensional map with many data points, **K-Nearest Neighbors** (KNN) [34] Search enables targeting K closes data points to the target query. It does this by measuring the distance from the query to every single other dot on the map. While this guarantees finding the actual closest neighbors, it is computationally expensive if there are a substansial amount of data points in the vector space.

Approximate Nearest Neighbor (ANN) [35] Search depict a set of search algorithms that offer a faster, though slightly less precise approach. ANN techniques work by building special data structures or using clever hashing methods to quickly narrow down the search to a smaller set. The **Ball-Boundary Decomposition (BBD) tree** [36] is a hierarchical data structure (tree) which organizes data points within a nested structure of ball-shaped regions. By navigating down the tree, the algorithm can quickly eliminate large portions of the data that are guaranteed not to contain the nearest neighbors. **Hierarchical Navigable Small World (HNSW)** [37] builds a multi-layered structure inspired from skip-grams [38], where the layers are stacked on top of each other. The top layers are very sparse, with only a few dots and long-range connections, allowing quick jumps to a general area of interest. The searches are initiated at the very top layer, then moving down through the layers, getting more and more precise until it reaches the bottom layer and finds the closest neighbors. **Locality Sensitive Hashing (LSH)** [39] maps similar data points to the same hash value. These hash values are then used to group data points into buckets. The search for nearest neighbors is restricted to the data points residing in the same buckets as the query, achieving sub-linear query time complexity. **Product Quantization** [40] compresses high-dimensional vectors by splitting them into lower-dimensional sub-vectors, which are then quantized by assigning them to the nearest centroid in a learned codebook for that subspace. **Faiss** [41] is a specialized library that provides indexing techniques, including implementations of ANN search methods such as HNSW, LSH, and Product Quantization.

LLMs have limited context windows, making it difficult to personalize responses based on past conversations. **MemoryBank** [22] tackles this issue by introducing a new memory mechanism for LLMs, which includes storage for interactions, event summaries, and user personality understanding. They implemented a dual-tower dense retrieval model similar to Dense Passage Retrieval [42] which ensures relevant recall, while an updater inspired by the Ebbinghaus Forgetting Curve [43] manages selective forgetting based on time and importance. **RET-LLM** [23] introduces a framework for LLMs, that extracts, stores, and recalls knowledge from text in the form of triplets, inspired by Davidsonian semantics. The framework comprises an LLM, a controller that manages the interaction flow, and the memory unit, which stores triplets and their vector representations for both exact and fuzzy retrieval using LSH.

ChatDB [24] introduces a framework for augmenting LLMs with external symbolic memory to enhance complex multi-hop reasoning. The core concept involves utilizing SQL databases as they enable precise recording, modification, querying, and deletion of data. It also employs a chain-of-memory approach, decomposing user input into a sequence of intermediate operation steps, each involving one or more SQL statements. **Spider 2.0** [25] is a new, challenging evaluation framework for text-to-SQL models, designed to reflect the complexities of real-world enterprise database use cases. Unlike previous benchmarks that often use simpler databases, Spider 2.0 features 632 problems derived from real-world applications with databases that can contain over 1,000 columns and are hosted on diverse systems like BigQuery [44] and Snowflake [45]. Solving these problems requires understanding not just the database schema but also dialect-specific documentation and even project-level codebases.

BIRD (BIg bench for laRge-scale Database grounded in text-to-SQL tasks) [46] is a new benchmark designed to bridge the gap between academic text-to-SQL research and real-world applications. Unlike prevalent benchmarks that focus primarily on database schema with limited data, BIRD features 12,751 text-to-SQL pairs across 95 large-scale databases (33.4 GB) spanning 37 professional domains. This emphasis on substantial database values introduces new challenges such as handling noisy data, grounding natural language questions in database values using external knowledge, and considering SQL query efficiency.

IV. Tool Use

Humans have been using tools for millennia. Bringing AI development closer to AGI requires making LLMs capable of using digital tools in order to achieve innovative solutions. This section explores papers that introduce unique ways to adapt tool use, replicating human behaviour.

LLMs have limitations to acquire external knowledge, particularly in real-time interaction with tools via APIs. **Toolformer** [11] create a few handwritten examples of how APIs should be used. The examples are pairs of LLM responses and text positions of where the API call should be placed during next-token prediction, letting the LLM train in a self-supervising way.

Some AI models struggle to process natural language as input, while others lack the ability to handle vision or speech. Additionally, complex tasks are often composed of multiple sub-tasks, which require a multi-model collaboration. **HuggingGPT** [12] uses LLMs as a central controller and as an orchestrator of a diverse set of AI tools from Hugging Face. The LLM performs task planning, by breaking down user requests into sub-tasks. For each sub-task it invokes an appropriate model based on their description. The LLM collects the chain of responses generated by the external models and creates a coherent output.

Another challenge lies in handling the large scale and constant updates of available APIs. **Gorilla** [13] tackles the issue of using models from Hugging Face and Pytorch Hub that host millions of constantly changing APIs, which cannot be described into a single context. Gorilla provides a fine-tuned LLaMA-based model combined with a Retriever-Aware Training (RAT) technique. At inference, the user's natural language prompt is optionally augmented with relevant API documentation retrieved using methods like BM25 or GPT-Index. They created a benchmark of 1600 ML APIs called APIBench, and proposed Abstract Syntax Tree (AST) to measure hallucinations.

ToolLLM [20] and **ToolAlpaca** [14] address the challenge of creating high-quality training data to enhance the external tool-use capabilities of LLMs. Instead of relying heavily on manual annotation, both papers leverage the capabilities of LLMs to automatically construct large datasets of tool-use instructions and corresponding solutions.

ToolAlpaca's [14] solution is a framework that automatically generates a tool-use corpus through a multi-agent simulation environment. The approach involves first constructing a diverse toolset by having LLMs generate structured documentation for over 400 real-world tool APIs across 50 categories, based on only brief introductions. Subsequently, a simulation environment is created with three distinct LLM-powered agents: a user, an assistant, and a tool executor. This process resulted in a corpus of 3,938 instances. Finally, this automatically generated corpus is used to fine-tune compact language models ToolAlpaca-7B and ToolAlpaca-13B.

ToolQA [16] and **API-Bank** [15] are benchmarks created to assess how well LLMs use external tools, beyond reliance on pre-existing knowledge. ToolQA centers on question answering using specific tools. Conversely, API-Bank examines the harder task of planning and using diverse APIs through dialogues.

ToolQA [16] addresses the limitations of existing benchmarks by ensuring that questions cannot be answered solely from the LLMs' pre-training data. The dataset comprises questions from 8 domains and defines 13 specialized tools for interacting with external knowledge sources like text, tables, and graphs. ToolQA is curated through an automated three-phase process involving reference data collection, human-guided question generation using templates, and programmatic answer generation to ensure accuracy.

API-Bank [15] addresses the critical need for a comprehensive benchmark to evaluate the capabilities of tool-augmented LLMs in planning, retrieving, and calling APIs. The authors developed API-Bank, with a runnable evaluation system with 73 diverse APIs. To facilitate the training of higher-quality tool-augmented LLMs, they introduce a large-scale training dataset that comprise of 1,888 dialogues and 2,138 APIs across 1,000 domains, generated cost-effectively using a system like ToolAlpaca. To validate the benchmark and training data, the authors fine-tuned Alpaca-7B on their dataset, resulting in Lynx. While experimental evaluations on API-Bank demonstrate that there remains room for improvement, Lynx significantly outperforms Alpaca in API utilization, establishing it as a valuable resource for future research aimed at enhancing the tool-use capabilities of LLMs.

RestGPT [47], like HuggingGPT, enables LLMs to perform complex tasks using external resources. However, RestGPT specifically focuses on interacting with real-world applications through RESTful APIs. It uses a step-by-step planning process to break down instructions into API calls. The system selects the right APIs, executes them, and understands the responses. This allows LLMs to effectively utilize a wide range of existing web services.

PLUTO (Planning, Learning, and Understanding for TOols) [17] introduces two main paradigms: "Plan-and-Retrieve" and "Edit-and-Ground". The P&R approach involves a neural retriever to shortlist relevant tools and an LLM-based planner that breaks down complex queries into actionable steps. The E&G approach uses LLMs to enhance tool descriptions based on how users actually employ them.

WebShop [31] is a simulated e-commerce website environment. It features 1.18 million real-world products scraped from Amazon and 12,087 crowd-sourced text instructions. WebShop requires an agent to navigate various webpages, formulate searches, understand product details, and select options to purchase a product matching the given specifications. The environment presents challenges in comprehending creative language, reformulating queries, processing noisy web text, and performing exploration across multiple pages. The authors collected over 1,600 human demonstrations and trained agents using reinforcement learning, imitation learning, and pre-trained models, achieving a best task success rate of 29%, significantly lower than human experts.

MRKL [48] and **OpenAGI** [49] incorporate various expert systems such as knowledge bases and planners to access domain-specific information.

V. Use Case Suitability

The diverse capabilities and architectural paradigms of agentic LLM systems make them optimally suited for different task domains and operational requirements. Broadly, these systems can be categorized into three functional focuses: tool-use, planning, and memory, each aligning with distinct real-world needs.

Tool-use focused systems like Toolformer [11], HuggingGPT [12], Gorilla [13], ToolAlpaca [14], API-Bank [15], ToolQA [16], and PLUTO [17] are best suited for environments where external integration and augmentation are necessary for task completion. For instance, [11] and [14] excel in scenarios requiring generalized or self-supervised API interaction, making them ideal for knowledge retrieval, fact-checking, or customer service chatbots. [12] is particularly well-aligned with multimodal tasks involving audio, image, or translation pipelines, such as in accessibility tools or creative applications. [13] stands out for developer-facing use cases and AI automation, given its retriever-aware API call generation capabilities.

When evaluation precision and real-world API usage fidelity are crucial, such as in healthcare, scheduling, or smart home management, the API-Bank benchmark [15] provides a comprehensive tool-use framework. [16] and [17], with their emphasis on reasoning and execution in real-time environments (e.g., databases, interpreters), are suited for domains like software diagnostics.

Planning-centric systems, including ReAct [18], FuRePa [19], DFSDT [20], and CoT [21], cater to use cases that demand complex reasoning, multi-hop decision-making, and real-time adaptation. ReAct [18] is particularly useful in interactive agent scenarios, such as virtual shopping assistants or robotics, where the system must reason and act iteratively. FuRePa [19] and DFSDT [20] are optimal for knowledge-intensive academic domains, supporting tasks like long-form question answering or research assistance due to their enhanced plan assessment and backtracking logic. Meanwhile, CoT [21] is especially effective in education and tutoring contexts, where step-by-step reasoning improves interpretability and learning outcomes.

Memory-enabled systems, such as MemoryBank [22], RET-LLM [23], ChatDB [24], and Spider 2.0 [25], address the need for persistence, personalization, and complex data manipulation. These systems are necessary in applications requiring long-term contextual retention or symbolic memory, such as personalized mental health assistants (MemoryBank [22]), enterprise knowledge support systems (RET-LLM [23]), or small business management tools (ChatDB [24]). Spider 2.0 [25] further extends memory capabilities to support multi-source SQL-based reasoning at an enterprise scale, facilitating data analytics and business intelligence over large-scale databases.

VI. Conclusion

Agentic AI has the potential to bridge the gap between the impressive language capabilities of LLMs and the demands of real-world problem-solving. This paper has explored the evolution of agentic workflows, surveyed key methodologies, and highlighted their transformative impact across industries. Our future work will focus on refining these architectures to integrate with custom systems through API compositions.

References

[1] Y. Dan, Z. Lei, Y. Gu, Y. Li, J. Yin, J. Lin, L. Ye, Z. Tie, Y. Zhou, Y. Wang, *et al.*, "Educhat: A large-scale language model-based chatbot system for intelligent education," *arXiv preprint arXiv:2308.02773*, 2023.

[2] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, "Chatlaw: Open-source legal large language model with integrated external knowledge bases," *CoRR*, 2023.

[3] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22199–22213, 2022.

[4] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, "Chemcrow: Augmenting large-language models with chemistry tools," *arXiv preprint arXiv:2304.05376*, 2023.

[5] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[7] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[8] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, *et al.*, "The rise and potential of large language model based agents: A survey," *Science China Information Sciences*, vol. 68, no. 2, p. 121101, 2025.

[9] Y. Li, H. Wen, W. Wang, X. Li, Y. Yuan, G. Liu, J. Liu, W. Xu, X. Wang, Y. Sun, *et al.*, "Personal llm agents: Insights and survey about the capability, efficiency and security," *arXiv preprint arXiv:2401.05459*, 2024.

[10] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, "A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges," *Vicinagearth*, vol. 1, no. 1, p. 9, 2024.

[11] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68539–68551, 2023.

[12] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *Advances in Neural Information Processing Systems*, vol. 36, pp. 38154–38180, 2023.

[13] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, "Gorilla: Large language model connected with massive apis," *Advances in Neural Information Processing Systems*, vol. 37, pp. 126544–126565, 2024.

[14] Q. Tang, Z. Deng, H. Lin, X. Han, Q. Liang, B. Cao, and L. Sun, "Toolalpaca: Generalized tool learning for language models with 3000 simulated cases," *arXiv preprint arXiv:2306.05301*, 2023.

[15] M. Li, Y. Zhao, B. Yu, F. Song, H. Li, H. Yu, Z. Li, F. Huang, and Y. Li, "Api-bank: A comprehensive benchmark for tool-augmented llms," *arXiv preprint arXiv:2304.08244*, 2023.

[16] Y. Zhuang, Y. Yu, K. Wang, H. Sun, and C. Zhang, "Toolqa: A dataset for llm question answering with external tools," *Advances in Neural Information Processing Systems*, vol. 36, pp. 50117–50143, 2023.

[17] T. Huang, D. Jung, and M. Chen, "Planning and editing what you retrieve for enhanced tool learning," *arXiv preprint arXiv:2404.00450*, 2024.

[18] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[19] Y. Zhu, Z. Luo, and G. Cheng, "Furthest reasoning with plan assessment: Stable reasoning path with retrieval-augmented large language models," *arXiv preprint arXiv:2309.12767*, 2023.

[20] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, *et al.*, "Toolllm: Facilitating large language models to master 16000+ real-world apis," *arXiv preprint arXiv:2307.16789*, 2023.

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[22] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "Memorybank: Enhancing large language models with long-term memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 19724–19731, 2024.

[23] A. Modarressi, A. Imani, M. Fayyaz, and H. Schütze, "Ret-llm: Towards a general read-write memory for large language models," *arXiv preprint arXiv:2305.14322*, 2023.

[24] C. Hu, J. Fu, C. Du, S. Luo, J. Zhao, and H. Zhao, "Chatdb: Augmenting llms with databases as their symbolic memory," *arXiv preprint arXiv:2306.03901*, 2023.

[25] F. Lei, J. Chen, Y. Ye, R. Cao, D. Shin, H. Su, Z. Suo, H. Gao, W. Hu, P. Yin, *et al.*, "Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows," *arXiv preprint arXiv:2411.07763*, 2024.

[26] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[27] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in neural information processing systems*, vol. 36, pp. 11809–11822, 2023.

[28] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, *et al.*, "Graph of thoughts: Solving elaborate problems with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 17682–17690, 2024.

[29] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8634–8652, 2023.

[30] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," *arXiv preprint arXiv:2010.03768*, 2020.

[31] S. Yao, H. Chen, J. Yang, and K. Narasimhan, "Webshop: Towards scalable real-world web interaction with grounded language agents," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20744–20757, 2022.

[32] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint arXiv:1809.09600*, 2018.

[33] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: a large-scale dataset for fact extraction and verification," *arXiv preprint arXiv:1803.05355*, 2018.

[34] V. Garcia, E. Debreuve, and M. Barlaud, "Fast k nearest neighbor search using gpu," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6, IEEE, 2008.

[35] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 604–613, 1998.

[36] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 891–923, 1998.

[37] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *CoRR*, vol. abs/1603.09320, 2016.

[38] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling.," in *LREC*, vol. 6, pp. 1222–1225, 2006.

[39] O. Jafari, P. Maurya, P. Nagarkar, K. M. Islam, and C. Crushev, "A survey on locality sensitive hashing algorithms and their applications," *arXiv preprint arXiv:2102.08942*, 2021.

[40] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 117–128, 2010.

[41] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," *arXiv preprint arXiv:2401.08281*, 2024.

[42] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering.," in *EMNLP (1)*, pp. 6769–6781, 2020.

[43] J. M. Murre and J. Dros, "Replication and analysis of ebbinghaus' forgetting curve," *PloS one*, vol. 10, no. 7, p. e0120644, 2015.

[44] S. Fernandes and J. Bernardino, "What is bigquery?," in *Proceedings of the 19th International Database Engineering & Applications Symposium*, pp. 202–203, 2015.

[45] B. Dageville, T. Cruanes, M. Zukowski, V. Antonov, A. Avanes, J. Bock, J. Claybaugh, D. Engovatov, M. Hentschel, J. Huang, *et al.*, "The snowflake elastic data warehouse," in *Proceedings of the 2016 International Conference on Management of Data*, pp. 215–226, 2016.

[46] 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. 42330–42357, 2023.

[47] Y. Song, W. Xiong, D. Zhu, W. Wu, H. Qian, M. Song, H. Huang, C. Li, K. Wang, R. Yao, *et al.*, "Restgpt: Connecting large language models with real-world restful apis," *arXiv preprint arXiv:2306.06624*, 2023.

[48] E. Karpas, O. Abend, Y. Belinkov, B. Lenz, O. Lieber, N. Ratner, Y. Shoham, H. Bata, Y. Levine, K. Leyton-Brown, *et al.*, "Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning," *arXiv preprint arXiv:2205.00445*, 2022.

[49] Y. Ge, W. Hua, K. Mei, J. Tan, S. Xu, Z. Li, Y. Zhang, *et al.*, "Openagi: When llm meets domain experts," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5539–5568, 2023.

A Review of Cross-Layer Security Vulnerabilities in IoT Communication Protocols

1st Berfin Ebrar Atabey
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University
Skopje, North Macedonia

2nd Dr. Sasho Gramatikov
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University
Skopje, North Macedonia

***Abstract*—The Internet of Things (IoT) has enabled smooth communication between devices across various domains, such as healthcare, smart homes, and industrial automation. However, the constrained nature of IoT devices and the lightweight design of communication protocols introduce significant security challenges. This paper presents a state-of-the-art review of the literature that focuses on the vulnerabilities of widely used IoT communication protocols, including MQTT, CoAP, AMQP, and HTTP, across different layers of the IoT architecture. The review analyzes protocol-specific security issues, common attack vectors against IoT protocols, and defense mechanisms. The findings reveal that most protocols prioritize performance and efficiency over security, leading to weak authentication, lack of encryption, and poor key management. Furthermore, the study highlights the importance of cross-layer security strategies and the need for lightweight mitigation techniques suitable for resource-constrained IoT environments.**

***Index Terms*—Internet of Things, IoT Protocols, IoT Security, IoT Protocols Security, Cross-Layer Security, IoT Layer Security.**

I. Introduction

The Internet of Things (IoT) has expanded rapidly in industries such as healthcare, smart homes, transportation, and industrial automation, enabling smooth data exchange and automation [1]. However, the growing number of interconnected IoT devices also introduce significant security challenges due to their constrained resources, limited power, and minimal storage capacities [2]. These limitations prevent the implementation of robust cryptographic protections, leaving IoT systems vulnerable to various cyber threats, including denial-of-service (DoS) attacks, data interception, and unauthorized device control [3]. IoT ecosystems rely on communication protocols like MQTT, CoAP, AMQP, and HTTP to ensure efficient data transmission between devices, gateways, and cloud services [4]. Many of these protocols prioritize performance and lightweight operation, often lacking adequate built-in security mechanisms [5]. Common vulnerabilities include weak or missing authentication, insufficient encryption, poor key management, and protocol misconfigurations. These weaknesses make IoT networks prime targets for attacks such as man-in-the-middle (MitM), device impersonation, and data exfiltration [6]. Moreover, applying strong security measures like encryption and certificate-based authentication introduces performance overhead, which resource-constrained IoT devices struggle to handle [7]. As a result, developers frequently face trade-offs between security and performance, leading to suboptimal implementations [8]. This paper presents a state-of-the-art review of layered IoT protocol security, focusing on vulnerabilities across different layers of IoT architectures. The goal is to provide a comprehensive understanding of existing security challenges in IoT communication protocols and identify areas requiring further research and improvement.

II. Methodology

The methodology applied during the research was a state-of-the-art review that analyzed existing research on IoT communication protocol vulnerabilities and related defense mechanisms.

A. Research Questions

1) What is the current state of identified vulnerabilities in IoT communication protocols across different layers of the IoT architecture?
2) What defense mechanisms and mitigation strategies have been proposed in the literature to address IoT vulnerabilities, particularly in resource-constrained IoT environments?

B. Paper Search and Selection Criteria

Relevant scientific papers were selected from peer-reviewed journals, conference proceedings, and academic databases such as IEEE Xplore, Google Scholar, ScienceDirect, and MDPI. The focus was placed on studies addressing security challenges across different layers of the IoT architecture, emphasizing application-layer protocols like MQTT, CoAP, and HTTP. The selected literature was reviewed, categorized, and compared to assess its relevance and suitability for inclusion in the study. After applying the selection criteria, a total of 31 peer-reviewed papers were chosen to proceed to the next stage, where inclusion and exclusion criteria were applied.

C. Inclusion and Exclusion Criteria

To ensure the relevance and quality of the reviewed literature, a set of inclusion and exclusion criteria was applied. Each paper was carefully reviewed based on the following criteria:

1) Inclusion Criteria:

- Studies published in peer-reviewed journals or conference proceedings.
- Papers focusing on IoT communication protocols and their security vulnerabilities.
- Research discussing cross-layer security issues or protocol-level mitigation strategies in IoT environments.
- Publications published within the last 8 years (2017–2024) to ensure relevance.

2) Exclusion Criteria:

- Papers not directly related to IoT protocol security or focusing solely on general IoT applications without addressing communication or security aspects.
- Non-peer-reviewed articles, blogs, white papers, or opinion pieces.
- Duplicate studies or papers lacking technical depth and methodology clarity.

D. Literature Categorization Approach

To ensure a structured review, the selected papers were categorized according to their primary focus areas. The groups include;

1) General protocol comparisons and security vulnerabilities.
2) Protocol-specific security issues such as MQTT and CoAP vulnerabilities.
3) Attack vectors and defense mechanisms, including DDoS and penetration testing.
4) Cross-layer and layered security perspectives.

Finally, nine papers were selected for the state-of-the-art review.

III. Literature Review

This section reviews selected studies that examine the security challenges and risks associated with widely used IoT communication protocols such as MQTT, CoAP, AMQP, and HTTP. The literature is organized based on protocol vulnerabilities, attack vectors, and proposed defense mechanisms, providing a comprehensive view of the current state of research in this field.

Naik [4] conducted a comparative analysis approach research where four widely used communication protocols—MQTT, CoAP, AMQP, and HTTP— were reviewed. The comparison criteria was based on these criterias, architecture, message size, QoS, security, power consumption, latency, and interoperability. Results of the paper showed that MQTT is a lightweight, low header size protocol that is best for constrained devices but it has limited interoperability and weakest security due to simple authentication and a lack of built-in encryption. So, MQTT was designed for constrained devices and networks, prioritizing low overhead and efficiency, not security. CoAP is found to have low overhead and the best power efficiency, uses UDP for lower latency. It also has moderate security via optional DTLS for message encryption, integrity, and authentication. Although constrained IoT devices may not fully implement them due to limited processing power and memory. AMQP has rich features (queueing, transactions), good security with TLS providing strong encryption and protecting data in transit, and it has a strong authentication mechanism. AMQP is observed to be designed for complex and secure business messaging systems with an offer of advanced security controls. These rich security features come with the cost of higher overhead, increased power consumption, and complexity. HTTP was found to be a global standard and has excellent interoperability, but heavyweight with the highest bandwidth, latency, and resource consumption. So not ideal for IoT. Also, it lacks native security features and depends on HTTPS (TLS/SSL) for secure communications. Overall, the conclusion of the paper shows that no single protocol fits all IoT use cases.

The paper by Tournier et al. [2] is a survey-based research to existing studies that analyze comparative analyses of seven IoT protocols (OS4I, BLE, Zigbee, Z-Wave, WirelessHart, LoRaWAN, SigFox). It introduces a generic IoT protocol stack model (5 layers: PHY, Data Link, Network, Transport, Application) to create a unified comparison framework. The comparison criteria include range, openness, interoperability, topology, and security practices. The authors categorize vulnerabilities and attacks into three layers;

- Packet-level attacks: e.g., passive/active cryptographic attacks
- Protocol-level attacks: e.g., MITM, Sybil, Flooding
- System-level attacks: e.g., Sinkhole, Selective forwarding

The results revealed that common vulnerabilities exist across IoT protocols, regardless of their design or layer. Many protocols lack security by design, especially proprietary or half-open ones like Z-Wave and SigFox, which offer limited visibility into their security mechanisms. Resource constraints further weaken these protocols, leading to poor implementation of encryption and authentication. Mesh topologies, such as in Zigbee and WirelessHart, improve resilience but increase exposure to attacks. Critical challenges were identified in Z-Wave and LoRaWAN due to weak or missing encryption and poor key management, while proprietary systems limit third-party security improvements and assessments. The overall conclusion of the paper is that there is no perfect IoT protocol; between efficiency, range, power consumption, and security there have to be trade-offs depending on the requirements. Cross–layer security is critical because vulnerabilities exist at different protocol layers.

Mansour et al. [9] conducted a comprehensive literature review comparing wireless technologies such as Wi-Fi, Bluetooth, ZigBee, and LoRaWAN, along with communication protocols including MQTT, CoAP, AMQP, and HTTP. Their analysis considers multiple layers of the IoT architecture: the Perception Layer, Network Layer, Middleware Layer, and Application Layer. The results presented in the paper indicate that security challenges in IoT systems vary across different layers. Therefore, cross-layer security approaches are critical

for ensuring the protection of IoT environments. The lack of standardization among protocols contributes to increased complexity and additional security challenges. IoT communication heavily relies on lightweight protocols such as MQTT and CoAP, which are designed for low-power devices but offer limited built-in security features. The paper also highlights that resource constraints in IoT devices prevent the implementation of strong cryptographic mechanisms, increasing their vulnerability. The overall conclusion is that some of the biggest security challenges IoT faces are heterogeneity, lack of standardization, and resource limitations. Protocol design trade-offs between efficiency and security expose IoT systems to cyber-attacks.

Mishra et al. [10] presented a performance evolution of four popular open-source MQTT brokers: Mosquitto, EMQ X, VerneMQ, and HiveMQ. They performed a stress testing by creating a custom-developed environment that simulated publishers and subscribers. The test measured latency, throughput, CPU usage, and memory consumption under varying loads like 1,000 to 30,000 simultaneous clients. Eventually, the brokers were evaluated according to their scalability, efficiency, and resource utilization. The findings indicate that in terms of latency and throughput, EMQ X and HiveMQ performed better under high loads. Mosquitto was the best-performing broker with small loads but struggled heavily under intense client connections. Mosquitto and Bevywise brokers, despite not being able to automatically scale, performed better than other brokers in resource-constrained environments. VerneMQ's performance was recorded as moderate, but stability was broken under extreme stress. ActiveMQ, with great scalability, showed excelling performance in distributed multi-core environments. The paper confirms that every broker shows different performances in terms of scalability and resource consumption.

Hintaw et al. [11] paper focuses on a survey-based study that performs a comprehensive analysis of attack vectors targeting different stages of MQTT protocol within IoT systems. Finally, the analysis of existing mitigation techniques and proposed solutions for each identified attack type is presented. The research explains that MQTT is widely used for its lightweight nature, but it is highly vulnerable to various attacks because of its lack of built-in security.

Key attack vectors are identified as;

- Authentication-based attacks.
- Authorization attacks.
- Communication attacks.
- Message attacks.
- Broker attacks.

The root causes of these vulnerabilities are MQTT prioritizing simplicity and a low overhead over security. Weak or optional authentication mechanisms. No native message encryption or verification. It relies on external TLS/SSL.

Presented solutions and defenses in the paper:

- Enforcing TLS/SSL encryption.
- Using token-based authentication instead of just username and password.
- Using token-based authentication instead of just username and password.
- Deploying intrusion detection systems to monitor MQTT traffic.

The general conclusion suggests that MQTT will remain vulnerable unless additional security layers are implemented. Developers designing MQTT systems should implement strong authentication, encryption, and access control mechanisms.

In the article, Dandotiya et al. [12] analyze CoAP protocol vulnerabilities and propose an enhanced security framework to improve its resistance to cyberattacks in IoT systems. The proposed model, generated considering attacks like eavesdropping, replay attacks, spoofing, and message injection, uses Datagram Transport Layer Security (DTLS) and AES-128 encryption for data transmission strengthening. The paper also provides a theoretical evaluation of how enhancements improve security without impacting the performance severely. The results showed that the CoAP protocol lacks native security mechanisms and is vulnerable to attacks in IoT environments. The proposed solutions are the integration of DTLS for securing communication at the transport layer, implementation of AES-128, encryption for ensuring data confidentiality, and replay attack prevention by including unique tokens or sequence numbers in CoAP to detect and prevent the attack.

Sanjaya et al. [13] conducted a research where a practical implementation and comparison of HTTP/2 and MQTT protocols for IoT-based document locker systems in government offices. A microcontroller-based prototype was developed and tested using both communication models. The comparison parameters were programming complexity, infrastructure needs, communication model, and resource overhead. The results showed that HTTP/2 with SSE offers simpler programming, reduced infrastructure, and efficient asynchronous communication suitable for single-client IoT systems. In contrast, MQTT provides a lightweight publish/subscribe model ideal for complex, multi-client IoT environments but introduces higher programming complexity due to broker configuration. While both protocols are effective, the study recommends HTTP/2 with SSE for applications where simplicity and minimal system overhead are prioritized. Contrary to that, MQTT is better suited for scalable systems that require many-to-many communication.

The paper from Lakshminarayana and Thilagam [14] presents a detailed analysis of how the new features of MQTT v5.0 can be exploited for Distributed Denial of Service (DDoS) attacks in IoT deployments. The authors conducted both theoretical modeling and experimental validation using a real-world IoT healthcare testbed. The authors developed specific attack scenarios exploiting three MQTT v5.0 features: CONNECT Properties, User Properties, and Flow Control. They designed two classes of DDoS attacks (Class I and

Class II) based on the attacker's capabilities (authenticated or not) and evaluated the resource impact (CPU, memory, latency) during these attacks. Class I attacks exploited the CONNECT Properties by flooding the broker with large or repetitive connection requests, leading to resource exhaustion. Class II attacks targeted the User Properties feature by sending metadata-heavy messages that consumed significant system resources and exploited the Flow Control mechanism by exceeding the permitted unacknowledged message limits, forcing disconnections. Experiments conducted on an IoT healthcare testbed demonstrated that such attacks could severely disrupt services, especially in critical applications like continuous patient monitoring, by significantly increasing resource usage and impacting system availability. The authors noted that while the experiments were limited to a healthcare testbed, broader evaluations are needed, and future research should focus on designing specific countermeasures to protect MQTT v5.0 deployments from these advanced attack vectors.

Mustafa et al. [15] conducted a comprehensive literature review of over 100 research articles. The review focuses on secure and energy-efficient cross-layer frameworks for IoT systems. The results proved that traditional security models are insufficient for dynamic IoT environments. They emphasized the importance of integrating advanced technologies such as AI-powered intrusion detection systems (IDS), blockchain, and quantum-safe cryptography to strengthen security while maintaining energy efficiency. Additionally, the authors stress the need for multi-layered security measures across the application, network, and sensor layers, recommending the use of lightweight cryptographic protocols and LPWAN technologies like LoRaWAN and NB-IoT for efficient long-range communication. The survey also identifies machine learning as a promising tool for dynamic threat detection in IoT networks. However, the paper notes that while these solutions are theoretically effective, further real-world implementation and testing are necessary to validate their practicality.

IV. Conclusion

This paper presented a state-of-the-art review of IoT communication protocols focusing on their security vulnerabilities across different layers of the architecture. The review revealed that protocols that are built for lightweight communication like MQTT and CoAP lack built-in security features. This makes them vulnerable to various attacks such as man-in-the-middle, replay, message injection, and DDoS. Advanced protocol versions, such as MQTT v5.0, introduce new features that improve functionality but also expand the attack surface when not properly secured. The findings highlight that resource constraints in IoT devices create a constant trade-off between implementing strong security measures and maintaining system performance. Protocols often rely on external security layers such as TLS/DTLS, which may not be fully utilized due to energy and processing limitations. In addition, the review emphasizes that cross-layer security strategies and AI-driven detection mechanisms are necessary to enhance the resilience of IoT systems, especially in critical applications like healthcare and industrial IoT. Future research should focus on developing lightweight, scalable, and protocol-specific security solutions, along with real-world validation of the proposed frameworks. As IoT deployments continue to grow, addressing these security gaps is essential to ensure the reliability, privacy, and resilience of IoT ecosystems.

References

[1] Y. Perwej, K. Haq, F. Parwej, M. M. Hassan, and M. Hassan, "The Internet of Things (IoT) and its application domains," International Journal of Computer Applications, vol. 182, no. 49, pp. 36–47, Apr. 2019, doi: 10.5120/ijca2019918887.

[2] J. Tournier, F. Lesueur, F. Le Mouël, L. Guyon, and H. Ben-Hassine, "A survey of IoT protocols and their security issues through the lens of a generic IoT stack," Internet of Things, vol. 16, p. 100264, 2021, doi: 10.1016/j.iot.2020.100264.

[3] M. A. Razzaque, M. Milojevic-Jevric, A. Palade, and S. Clarke, "Middleware for Internet of Things: A survey," IEEE Internet of Things Journal, vol. 3, no. 1, pp. 70–95, Feb. 2016, doi: 10.1109/JIOT.2015.2498900.

[4] N. Naik, "Choice of effective messaging protocols for IoT systems: MQTT, CoAP, AMQP and HTTP," 2017 IEEE International Systems Engineering Symposium (ISSE), Vienna, Austria, 2017, pp. 1–7, doi: 10.1109/SysEng.2017.8088251.

[5] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2347–2376, 2015, doi: 10.1109/COMST.2015.2444095.

[6] A. Munir, I. A. Sumra, R. Naveed, and M. A. Javed, "Techniques for authentication and defense strategies to mitigate IoT security risks," Journal of Computing & Biomedical Informatics, vol. 7, no. 1, pp. 377–388, Jun. 2024, doi: 10.56979/701/2024.

[7] S. Singh, P. K. Sharma, S. Y. Moon, and J. H. Park, "Advanced lightweight encryption algorithms for IoT devices: survey, challenges and solutions," Journal of Ambient Intelligence and Humanized Computing, vol. 15, no. 6, pp. 1625–1642, 2024, doi: 10.1007/s12652-017-0494-4.

[8] M. Rana, A. Shafiq, I. Altaf, M. Alazab, K. Mahmood, S. A. Chaudhry, and Y. B. Zikria, "A secure and lightweight authentication scheme for next generation IoT infrastructure," Computer Communications, vol. 165, pp. 85–96, 2021, doi: 10.1016/j.comcom.2020.11.002.

[9] M. Mansour, A. Gamal, A. I. Ahmed, L. A. Said, A. Elbaz, N. Herencsar, and A. Soltan, "Internet of Things: A Comprehensive Overview on Protocols, Architectures, Technologies, Simulation Tools, and Future Directions," Energies, vol. 16, no. 8, p. 3465, Apr. 2023, doi: 10.3390/en16083465.

[10] B. Mishra, B. Mishra, and A. Kertesz, "Stress-Testing MQTT Brokers: A Comparative Analysis of Performance Measurements," Energies, vol. 14, no. 18, p. 5817, Sep. 2021, doi: 10.3390/en14185817.

[11] A. J. Hintaw, S. Manickam, M. F. Aboalmaaly, and S. Karuppayah, "MQTT vulnerabilities, attack vectors and solutions in the Internet of Things (IoT)," IETE Journal of Research, vol. 69, no. 6, pp. 3368–3397, 2023, doi: 10.1080/03772063.2021.1912651.

[12] N. Dandotiya, A. S. Dandotiya, R. Dubey, S. Gupta, S. Sharma, and A. K. Sharma, "Enhancing CoAP for Secure IoT Communication," in Proceedings of the 2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N), Greater Noida, India, 2024, pp. 1167–1172, doi: 10.1109/ICAC2N63387.2024.10895113.

[13] R. Sanjaya, A. K. Pamudji, and E. Vania, "Comparative Analysis and Practical Implementation of IoT Communication Using HTTP/2 and MQTT," in Proceedings of the 2024 7th International Conference on Green Technology and Sustainable Development (GTSD), Ho Chi Minh City, Vietnam, 2024, pp. 221–226, doi: 10.1109/GTSD62346.2024.10674902.

[14] S. Lakshminarayana and P. S. Thilagam, "Next-Generation DDoS Attacks on IoT Deployments: Targeting the Advanced Features of MQTT v5.0 Protocol," IEEE Internet of Things Journal, early access, 2025, doi: 10.1109/JIOT.2025.3549784.

[15] R. Mustafa, N. I. Sarkar, M. Mohaghegh, and S. Pervez, "A Cross-Layer Secure and Energy-Efficient Framework for the Internet of Things: A Comprehensive Survey," Sensors, vol. 24, no. 22, p. 7209, Nov. 2024, doi: 10.3390/s24227209.

A State-of-the-Art Review on Social Media Exposure and Cyber Threats

Lydiah Wanjugu Muigwa
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje (UKIM)
Skopje, North Macedonia
lydiahmuigwa@gmail.com

Ivan Chorbev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje (UKIM)
Skopje, North Macedonia
ivan.chorbev@finki.ukim.mk

***Abstract*—Social media is now a massive part of how we connect personally and professionally, but it comes with real dangers. Whenever we share personal details online, we risk exposing ourselves to scams, identity theft, and manipulation. Even with growing privacy concerns and laws like GDPR and CCPA in place, many of us still find it hard to control our digital footprints. The settings are often confusing, and the tools available aren't always user-friendly. In this paper, we look closer at existing privacy tools, sorting them by how they work and how easy they are to use. We also evaluate how well they protect users from social media threats. Our research reveals significant gaps like the lack of ways to manage data across multiple platforms, personalized risk assessments, and better privacy education. To contextualize these findings, we present an integrated conceptual framework illustrating how digital footprints contribute to privacy risks and where current solutions fall short in mitigating them.**

***Index Terms*—Social media privacy, digital footprint, aggregation paradox, phishing, OSINT, cybersecurity education**

I. Introduction

Social media has completely changed how we interact, share, and shape our identities online. But these platforms thrive on visibility, engagement, and data sharing, which often comes at the cost of our privacy and security. Cybercriminals are using the personal details we post to create convincing phishing scams, steal identities, and manipulate people through carefully crafted social engineering attacks [1]. Despite privacy laws like GDPR and CCPA, staying safe online isn't easy. Privacy settings are scattered and hard to manage, and platform terms, often the sole mechanism for 'consent,' are illusory, as firms unilaterally modify them without notice [2]. [3]found that 60%

While AI-driven Open Source Intelligence (OSINT) tools can automate threat detection, [4], these solutions remain inaccessible to non-experts. There are not enough user-friendly tools to help people understand or control their exposure, as current solutions ignore the 'aggregation paradox' of cross-platform data integration.

In this paper, we deeply dive into the tools and strategies designed to protect personal data on social media. Some alert you after a breach happens, others tweak platform-specific settings, and a few are advanced tools for cybersecurity experts. We examine their shortcomings, whether they are poor usability, limited coverage, or a lack of built-in privacy education. Ultimately, we aim to map out where personal data protection stands today, pinpoint the most significant vulnerabilities, and highlight areas needing more attention so future solutions can empower users to stay safe online.

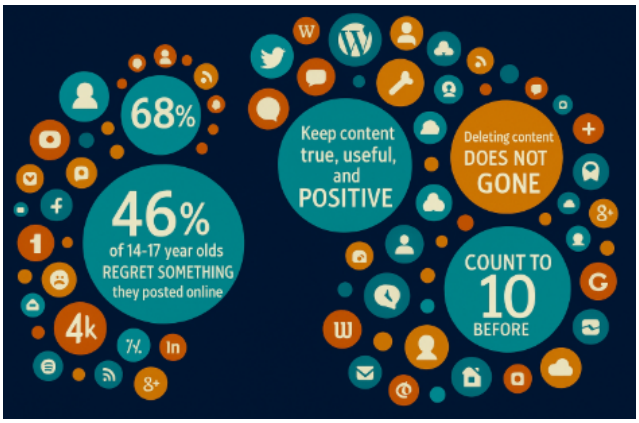


Fig. 1. Digital Footprints in Social Media

II. Background

Social media platforms have unintentionally become the most comprehensive surveillance tool ever created, not by governments but by ordinary users themselves. Each like, check-in, and comment contributes to a sprawling digital identity exceeding what any individual realizes they've revealed. This section examines how modern cyber attacks harvest and weaponize these self-created data repositories.

A. Digital Footprints and OSINT

A digital footprint is the collection of all data left behind by a person's activities in digital systems. Unlike a fixed log, it grows and changes over time as new data links to existing records, creating interconnected patterns. When users post about their new jobs on LinkedIn, check in at a restaurant on Facebook, and share vacation photos on Instagram, they create interconnected data points that OSINT tools can stitch into startlingly accurate profiles. Figure 1 shows some digital footprints in social media.

Malicious actors leverage these profiles to bypass traditional security measures entirely. AI-powered OSINT tools further enable attackers to automate identity reconstruction from fragmented data [4].

Why hack a password when you can reconstruct an identity from social posts? This threat is compounded by platforms' unilateral data-sharing clauses, which permit transfers to third parties during mergers [2], exposing users to unforeseen risks like the Google-Fitbit acquisition.

Recent studies demonstrate that such synthesized identities enable 83% of successful business email compromise attacks [5]. The rise of generative AI has rapidly increased this threat, allowing attackers to produce convincing deepfake audio using just minutes of publicly available voice clips.

B. Social Media as a Threat Actor

Each platform has developed its distinct vulnerability profile based on design choices and user behaviors:

LinkedIn's professional focus makes it ideal for organizational mapping. Attackers routinely use job transition announcements to identify newly promoted employees as prime targets for fake recruiter scams or credential theft. The platform's endorsement features further enable reputation hijacking, where criminals build fake profiles boosted by legitimate-looking connections.

Facebook's emphasis on personal connections has created a golden mine for relationship exploitation. The platform's "Friends You May Know" algorithm frequently reveals unexpected connections between targets, while memorialized accounts of deceased users become permanent vulnerabilities that family members can't secure.

Instagram's visual nature introduces physical security risks that users rarely consider. A 2024 penetration testing study found that 41% of corporate security breaches began with intelligence gathered from employee Instagram posts, including building layouts visible in background reflections and work badge details captured in selfies.

Cybersecurity researchers have identified a critical issue, the 'aggregation paradox' [6], which compounds these platform-specific weaknesses. This paradox refers to the deceptive nature of seemingly harmless individual data points. While each data point might seem harmless, the combinatorial power of cross-platform data integration creates catastrophic privacy failures. [7] found that location tags from Instagram posts and LinkedIn job titles enabled attackers to infer home addresses and routines, demonstrating how cross-platform data integration creates systemic risks.

This means that when seemingly harmless data points from different platforms are combined, they can reveal sensitive information. Neither current privacy regulations nor platform safeguards adequately address this systemic risk, as they focus on individual data elements rather than emergent patterns created by data convergence.

This background sets the stage for examining why current privacy tools fail to address these multidimensional threats. The urgency of this gap becomes increasingly apparent as AI-powered attack automation outpaces conventional defense mechanisms. The following section will analyze how existing solutions attempt and largely fail to mitigate these risks.

III. Privacy and OSINT Tools: A Survey of Existing Solutions

To understand the current landscape of privacy protection and data exposure mitigation, this section reviews various tools and frameworks, ranging from people search engines to advanced OSINT platforms. These tools differ in scope, user accessibility, depth of analysis, and cross-platform support.

We categorize the tools into four groups:

1) Reactive Alert Systems: Tools that notify users after a data breach or leak.
2) Platform-Specific Privacy Features: Built-in tools offered by social platforms for privacy review and settings control.
3) People Intelligence and Search Engines: Services for aggregating personal data from public records and social activity.
4) General-Purpose OSINT Tools: Comprehensive intelligence platforms for mapping digital footprints, often requiring technical expertise.

However, OSINT tools face systemic challenges, including jurisdictional conflicts and unreliable data aggregation [8]. For instance, profiling threats across social media requires cross-platform data integration, yet neither technical nor regulatory solutions address the 'aggregation paradox' (Section II.B).

A. Limitations of Current Tools for Everyday Users

OSINT tools, while powerful, are complex and designed for investigative professionals rather than everyday users [9]. A systematic review of 163 AI-OSINT tools found that 89% use binary risk classification (safe/unsafe) and fail to account for emergent AI-driven threats like deepfakes [4], leaving users vulnerable to evolving tactics. This aligns with [10] observation that nontechnical users struggle with privacy settings, leaving them vulnerable to threats like fake profiles and phishing.

Most platforms reviewed, including tools like Maltego, OSINT Framework, and Epieos, are optimized for cybersecurity analysts, journalists, and researchers. These tools prioritize the breadth and speed of data collection but offer minimal support for nontechnical users who want to understand and act on their digital exposure.

A significant problem is the lack of clear guidance on what to do next. Even if users discover their data exposure, platform terms rarely offer actionable remedies, partly because firms reserve the right to modify privacy policies retroactively [2]. These decisions require awareness and confidence, which most tools fail to build.

An illustrative case is a real report from OSINT Industries, a real-time search platform. The report mapped dozens of online profiles, exposed hints of phone numbers and breached credentials, and linked to public social media accounts. However, no advice about remediation, account security, or privacy

Tool	Type	Key Features	Limitations
Have I Been Pwned	Breach Alert	Checks emails against known breaches	Reactive only
Facebook Privacy Checkup	Platform Settings	Guides users through Facebook privacy settings	Platform-specific; no cross-platform support
Maltego	OSINT Tool	Maps relationships between emails, domains, and social data	Requires technical expertise
SpiderFoot	OSINT Automation	Aggregates data from 100+ sources	Requires technical expertise
OSINT Framework	OSINT Directory	Links to categorized tools for public data mining	Requires deep manual exploration
Spokeo	People Intelligence	Aggregates social and demographic data	Subscription-based
PimEyes	Face Recognition	Finds matching faces across public images	Potential privacy issues
Epieos	OSINT Search	Email-based people search and metadata analysis	Requires interpretation;
OSINT Industries	Real-Time OSINT	Live search with anonymized querying	Lacks integrated risk scoring or remediation options
Social Searcher	Social Monitoring	Tracks mentions across public social networks	Focused on brands; limited for individual privacy

TABLE I
OVERVIEW OF OSINT TOOLS FOR ASSESSING ONLINE EXPOSURE

settings was given. For everyday users, this leaves them with an overwhelming snapshot of their exposure but no clarity on how to reduce it.

Furthermore, many tools operate under a subscription model, adding significant financial barriers. While researchers may have institutional access or justification for these tools, individual users face usability and cost hurdles. This raises equity concerns about who gets to manage their digital footprint effectively.

These challenges point to a broader gap in privacy management, access to intuitive, actionable, and affordable tools that move beyond exposure to mitigation.

IV. Social Engineering and Phishing Tactics

Cybercriminals exploit human psychology and social context. Research shows that phishing messages that leverage social networks achieve a 72% success rate—4.5× higher than generic attacks [1]. This aligns with [5], who found that even trained individuals remain vulnerable to highly targeted spear phishing, with a 54% click rate after training. Their study highlights that individual-loss-framed training reduces susceptibility by 12%, suggesting personalized risk messaging could mitigate threats.

Threat actors gather personal information through social media. They look for:

- Job titles and workplace details
- Personal relationships and interests
- Recent posts or activities

With these details, they craft convincing messages, adding urgency or fake credibility to pressure their targets. For example, an email referencing a recent social media post is much harder to dismiss as spam. This shift from generic scams to highly personalized ”spear-phishing” makes these attacks especially dangerous [11].

The research underscores the critical need for systems that help users identify which pieces of their public data are most dangerous and how to mitigate their availability. This awareness is key to staying one step ahead of cyber attackers.

V. Privacy Education Effectiveness

Numerous studies affirm the importance of cybersecurity education in reducing user vulnerability. However, the format and delivery of these interventions play a significant role in their effectiveness.

- Interactive training outperforms passive methods, but format matters. [9] found that hands-on labs and gamified simulations had the highest user engagement and knowledge retention, yet only 16% of participants had experienced such training. This underscores the gap between proven methods and real-world adoption, a key challenge for privacy education.
- Personalized, context-specific lessons result in higher knowledge retention.
- Repeated engagement and real-time feedback help build long-term behavioral change.

Regrettably, the market lacks privacy tools that integrate effective learning modules. [7] found that users often feel ’overwhelmed’ by their digital traces, underscoring the need for tools that offer real-time guidance, such as nudges when posting high-risk data (e.g., location tags). Even when education is provided, it often falls short, being either generic or optional, thereby failing to influence user behavior [3] significantly.

While the limitations of current tools and privacy education programs are well documented, few studies present an integrated view of how digital footprints, OSINT-based exposure mapping, privacy risks, and educational interventions interrelate. As a synthesis of insights from prior work, we include a conceptual framework (Fig. 2) that visualizes this flow from personal data exposure to risk, highlights critical awareness gaps, and organizes reviewed mitigation strategies, including nudges and simulations.

VI. Critical Challenges and Future Research Directions

A. Critical Unresolved Challenges

The framework in Fig. 2 maps the fragmentation of current tools and educational interventions, emphasizing the need for holistic design thinking in future research. Based on this, our systematic review identifies persistent research and tooling gaps in social media privacy. Although current solutions

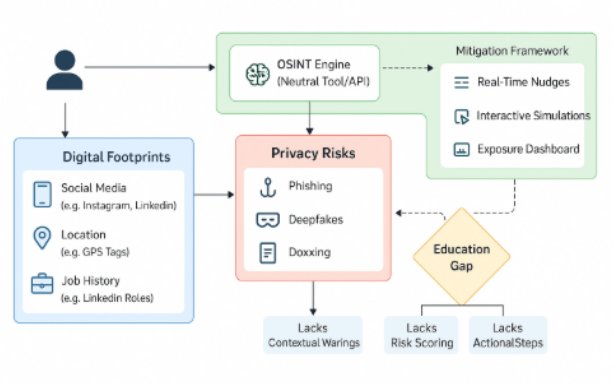


Fig. 2. Conceptual Framework Integrating Literature on Digital Footprints, OSINT Analysis, Privacy Risks, Education Gaps, and Mitigation Strategies.

address platform-specific risks, they largely fail to counter cross-platform threats enabled by data aggregation.

As [4] notes, only 11% of AI-OSINT tools support cross-platform correlation, worsening the aggregation paradox, where combined data fragments create unintended privacy risks.

Key shortcomings include:

- Decade-old posts remain accessible, yet tools rarely flag them.
- 89% of the tools use binary risk classification (safe/unsafe) rather than granular scoring.
- No tools account for emergent risks like deepfake profiling from public photos.
- Privacy modules lack:
  - Contextual triggers (for example, real-time warnings for high-risk posts).
  - Long-term impact: Studies show a $< 30\%$ behavioral change [IEEE SP 2024].
- Simple tools (e.g., browser privacy checkers) lack actionable insights.
- Advanced tools (e.g., OSINT frameworks) require technical expertise absent in 92 % of users [Eurostat, 2023].

B. High-Priority Research Vectors

To address these gaps, we prioritize three research directions:

Create tools that connect privacy risks across different platforms. Prioritize multilingual support and API interoperability, as [4] notes that these are critical for global threat intelligence but absent in 92% of current tools. Generate dynamic risk scores (e.g., "Your shared data enables eight attack variants"). A key challenge is overcoming the platform's API restrictions.

Integrate contextual privacy warnings during user posting (e.g., real-time alerts: Job title + hometown increases phishing risk). [5] demonstrates that loss-framed warnings ('You could lose X') outperform generic advice. This aligns with [9], who found that interactive training with instant feedback (e.g., CPR simulations) sustains behavioral change, a framework applicable to privacy education."

Create open-source tools that automate exposure mapping for non-experts. However, [8] cautions that such tools must balance usability with analytical depth, as overly simplistic dashboards (e.g., browser privacy checkers) lack actionable insights. Like their proposed 'Neighborhood Watch' framework, a hybrid approach could merge automation with guided remediation steps.

C. Implementation Road map

Translating these directions into practice requires:

- Industry-Academia Collaboration: Standardized APIs for privacy tool development and shared datasets of multi-platform exposures.
- Regulatory Support: Mandates for quantifying policy privacy risks (e.g., "This data type increases account hijacking risk by X
- User-Centered Design: Co-creation studies with nontechnical populations to ensure accessibility and adoption.

These priorities reflect a growing academic consensus: while awareness of digital risks is increasing, systemic shortcomings in usability, contextual feedback, and cross-platform integration persist. The conceptual framework offers a structured lens for researchers to assess where and how interventions could best support users moving forward.

References

[1] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Communications of the ACM*, vol. 50, no. 10, pp. 94–100, 2007.

[2] T. D. Haley, "Illusory privacy," *Ind. LJ*, vol. 98, p. 75, 2022.

[3] O. Pantelic, K. Jovic, and S. Krstovic, "Cookies implementation analysis and the impact on user privacy regarding gdpr and ccpa regulations," *Sustainability*, vol. 14, no. 9, p. 5015, 2022.

[4] T. O. Browne, M. Abedin, and M. J. M. Chowdhury, "A systematic review on research utilising artificial intelligence for open source intelligence (osint) applications," *International Journal of Information Security*, vol. 23, no. 4, pp. 2911–2938, 2024.

[5] A. Burns, M. E. Johnson, and D. D. Caputo, "Spear phishing in a barrel: Insights from a targeted phishing campaign," *Journal of Organizational Computing and Electronic Commerce*, vol. 29, no. 1, pp. 24–39, 2019.

[6] A. Zelenkauskaite, "Remediation, convergence, and big data: Conceptual limits of cross-platform social media," *Convergence*, vol. 23, no. 5, pp. 512–527, 2017.

[7] E. Nicol, J. Briggs, W. Moncur, A. Htait, D. P. Carey, L. Azzopardi, and B. Schafer, "Revealing cumulative risks in online personal information: a data narrative study," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–25, 2022.

[8] A. Yeboah-Ofori and A. Brimicombe, "Cyber intelligence & osint: Developing mitigation techniques against cybercrime threats on social media a systematic review july 2017," *International Journal of Cyber-Security and Digital Forensics*, vol. 7, no. 1, pp. 87–99, 2018.

[9] M. J. Dupuis and C. Gordon, "Evaluating prevalence, perceptions, and effectiveness of cyber security and privacy education, training, and awareness programs," in *The Colloquium for Information Systems Security Education. New Orleans, Louisiana, USA*, 2018.

[10] S. Ali, N. Islam, A. Rauf, I. U. Din, M. Guizani, and J. J. Rodrigues, "Privacy and security issues in online social networks," *Future Internet*, vol. 10, no. 12, p. 114, 2018.

[11] M. Bossetta, "The weaponization of social media: Spear phishing and cyberattacks on democracy," *Journal of international affairs*, vol. 71, no. 1.5, pp. 97–106, 2018.

Large Language Models for Predictive Neurology: A Neurogenomic and Neurogenetic Perspective

1st Djansel Bukovec
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
dzanselbukovec@gmail.com

2nd Blagojche Gjorgjioski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
gj.blagojce@gmail.com

Kostadin Mishev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
kostadin.mishev@finki.ukim.mk

Monika Simjanoska Misheva
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
monika.simjanoska@finki.ukim.mk

***Abstract*—Large Language Models (LLMs) are emerging as powerful tools for neurology, offering novel capabilities in clinical decision support for complex neurodegenerative diseases like Alzheimer's disease (AD). This paper outlines an integrated LLM approach for predictive neurology, focusing on diagnostics, biomarker discovery, and risk stratification. We propose a structured pipeline where generative LLMs generate hypotheses and discriminative LLMs validate these against multi-modal data. A case study on early AD risk assessment illustrates the workflow: an LLM-driven analysis of genomic and epigenomic patient data reclassifies ambiguous variants and identifies personalized intervention strategies.**

***Index Terms*—Large Language Models, LLMs, Generative AI, Discriminative AI, Alzheimer, Neurology, Neurogenetics, Neurogenomics, Neurodegenerative diseases.**

I. Introduction

Transformer-based LLMs have rapidly advanced natural language understanding in healthcare. In biomedical domains, models like BioBERT [1] and BioGPT [2] are pre-trained on scientific text to recognize biomedical entities and relations, enabling literature mining and question-answering. Generative AI methods are being used to draft clinical reports, summarize patient histories, and aid in diagnostic reasoning. LLMs can synthesize symptoms and test results into probable diagnoses in plain language, assisting clinicians with decision support. However, base LLMs can hallucinate or miss domain-specific details, motivating adaptations for specialized tasks.

The neurogenomic era brings high-throughput sequencing and multi-omics to neurology. Diseases like AD involve genomic variants (e.g. in *APP*, *PSEN1*/2), polygenic risk scores, transcriptomic changes, and epigenetic alterations. Making sense of this neurogenetic data is complex: one must interpret variants of uncertain significance (VUS), identify gene expression patterns, and correlate epigenetic markers with disease progression. Traditional approaches include databases (ClinVar, OMIM) and predictive algorithms for variant pathogenicity, as well as network and knowledge graph models to connect genes, pathways, and phenotypes. Knowledge graphs (KG) have been used for AD to integrate data and suggest mechanisms or drug targets. However, KGs require curated knowledge and may not capture novel associations not yet in databases.

LLMs can complement these methods by reading and summarizing vast unstructured biomedical literature and patient data. For instance, a specialized model like GeneGPT [3] augments an LLM with NCBI gene databases, allowing it to answer genomics questions with high accuracy. Such models can retrieve gene-specific information and interpret it in context (e.g., linking a genotype to potential functional impact). Other emerging efforts include LLMs trained on genomic sequences or epigenomic data (e.g., EpiGePT [4], a transformer for predicting epigenomic signals) to support interpretation of how genetic and epigenetic factors interplay in neurological diseases. By incorporating domain-specific knowledge, these LLMs have achieved state-of-the-art performance on benchmarks, illustrating their potential: GeneGPT [3], for example, scored 0.83 on a gene understanding benchmark versus 0.12 by a general model (ChatGPT). iLLMAC [5] shows strong potential for improving non-invasive diagnosis in neuro-oncology, especially for detecting and classifying primary brain tumors and CNS metastases. By analyzing cell-free DNA (cfDNA) from cerebrospinal fluid (CSF) or blood, it can recognize unique fragmentation patterns linked to cancers like gliomas, meningiomas, or brain metastases. Since iLLMAC is data-agnostic and works with various sequencing methods, its cfDNA end-motif analysis could be applied to other diseases involving cell death and distinct DNA patterns. It might also support liquid biopsy approaches for neurodegenerative

Funded 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.

diseases such as Alzheimer's or Parkinson's, where cfDNA could reflect neuron damage or epigenetic changes.

In summary, LLMs, especially when tailored to biomedical data, offer a new avenue for integrating genomic, clinical, and literature information into a cohesive decision support tool for neurology.

II. LLM-Supported Neurogenomic Decision Pipeline

This section outlines the proposed LLM-assisted pipeline for neurology clinical decision support, designed to integrate generative and discriminative reasoning for complex cases. Figure 1 provides an overview of the pipeline, which consists of the following components and steps:

Data Inputs: The system accepts multi-modal patient data, reflecting the multifaceted nature of neurological disorders. Inputs can include genomic data (e.g. gene panel or whole genome sequences with variant annotations), transcriptomic profiles (gene expression data), epigenomic patterns (DNA methylation, histone modification profiles), and relevant clinical information (phenotypes, family history, cognitive assessments). This comprehensive input ensures that the LLMs consider both molecular data, and clinical context in tandem.

Problem Definition: A clinical or research question is first defined based on the patient scenario. For example, the neurologist might ask: "Does this patient's *PSEN1* variant contribute to early-onset Alzheimer's, and what targeted therapy could mitigate risk?" or "Identify biomarkers from multi-omics data that predict rapid AD progression." A central coordinator module interprets this query and strategizes the analysis flow. In implementation, this could be a simple control script or logic that triggers the appropriate LLMs in sequence.

Hypothesis Generation: A generative LLM specialized in biomedical knowledge is invoked to address open-ended aspects of the query. For instance, a genomics-focused LLM (analogous to GeneGPT) could survey known literature and databases to generate hypotheses about a variant's effect or propose potential biomarkers. The model might output statements like "The *PSEN1 c.1234A>G* variant is predicted to disrupt amyloid precursor protein processing, possibly accelerating plaque formation," citing supporting evidence from prior studies. In biomarker discovery queries, it could suggest candidate biomarkers (genes or epigenetic marks) that warrant further examination. The generative LLM's strength lies in its ability to synthesize information and propose novel connections (even extrapolating from sparse data), improving the exploratory breadth of the analysis. Output would be a set of hypotheses or preliminary answers with justifications.

Analysis & Validation: Next, one or more discriminative LLMs are deployed to rigorously test the generative model's hypotheses against the data. For example, a classification-oriented LLM (analogous to iLLMAC [5]) processes the patient's genomic data to predict the pathogenicity of the suggested variant. It uses learned knowledge of variant pathogenicity (incorporating features like evolutionary conservation, known disease associations, etc.) to assign a probability

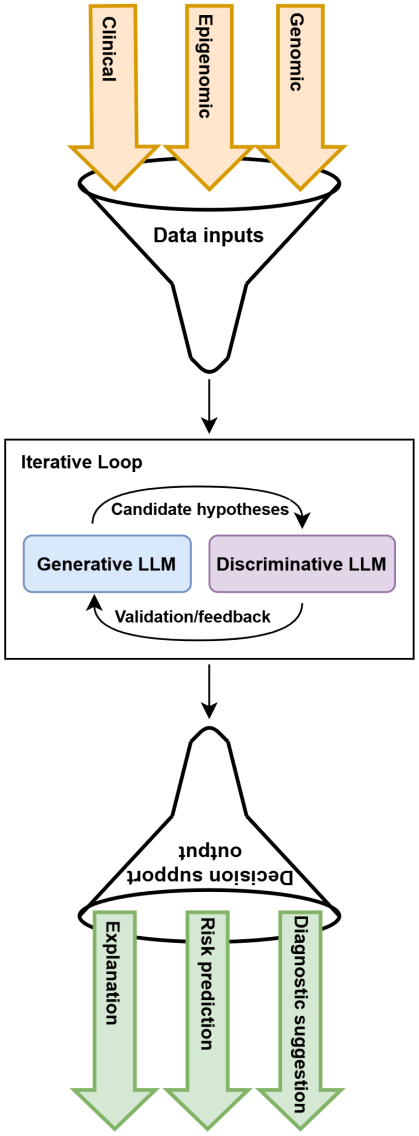


Fig. 1. LLM-based pipeline for generating and validating neurogenomic clinical insights.

or category (e.g. likely pathogenic vs benign) to the variant. Likewise, for biomarkers, a discriminative model might evaluate the statistical association between suggested markers and patient outcomes across available datasets. These models output results such as confidence scores and explanatory highlights (e.g. which gene features or clinical factors most influenced the prediction). By doing so, they validate or refute the generative model's outputs in a data-driven manner. The discriminative step grounds the analysis in evidence, reducing false positives from the generative stage.

Iterative Refinement: A key aspect of the pipeline is the potential for iteration. After initial validation, the system can

loop back: if the discriminative analysis finds a hypothesis only partially supported or identifies uncertainties, the generative model can be prompted to refine its hypothesis or generate an alternative. For instance, if one variant is deemed likely benign, the generative LLM might be asked "Given that Variant A is likely benign, what other explanations or variants could account for the family history?" – leading it to propose new hypotheses (perhaps implicating another gene or an environmental factor). The discriminative model would then test these. This feedback loop continues until a satisfactory consensus or list of top-ranked hypotheses is reached. Such iterative interplay ensures that the generative creativity is continually checked by discriminative rigor, improving result reliability.

Multi-Modal Integration: Throughout the above steps, specialized LLMs can focus on different data modalities for parallel insights. For example, an epigenomics-focused LLM (e.g. akin to EpiGePT) can analyze epigenetic input data independently to flag significant patterns (like "hypermethylation in promoters of neuroplasticity genes"). Meanwhile, a clinical note analysis LLM might summarize the patient's medical history to highlight any symptom patterns. In the pipeline, these modality-specific analyses are integrated – either by the coordinator or via the generative model compiling all evidence. The end result is a consolidated reasoning report that links genetic findings with epigenetic signals and clinical context. This step leverages the ability of multiple LLMs to work on different data but contribute to a unified outcome.

Output & Explanation: Finally, the system produces an output for the clinician or researcher. This includes the predicted answer or decision support (e.g. "Patient's variant is Likely Pathogenic and associated with early AD; aggressive preventive therapy is recommended") along with explanations: natural-language rationale, supportive references (literature, database entries the models accessed), and even uncertainty estimates. The goal is to provide a transparent, interpretable result to the end-user, to enhance trust and facilitate verification. Clinicians can review the supporting evidence, much like they would read an AI-generated report, and cross-check critical points.

The entire pipeline is designed to operate within a clinician's workflow. For example, it could be integrated into an electronic health record system: once genetic test results are uploaded, the pipeline auto-generates a report on variant significance and possible next steps. Importantly, Figure 1 illustrates this pipeline flow, from data input through hypothesis generation, validation, and integrated output. The figure highlights the roles of generative LLM (creative brainstorming) vs discriminative LLM (fact-checking and prediction), showing how their interplay yields a refined decision support result.

III. Case Study: Early Alzheimer's Disease Risk Assessment

To demonstrate the pipeline in action, we present a case scenario of an AD risk assessment. In this fictional but

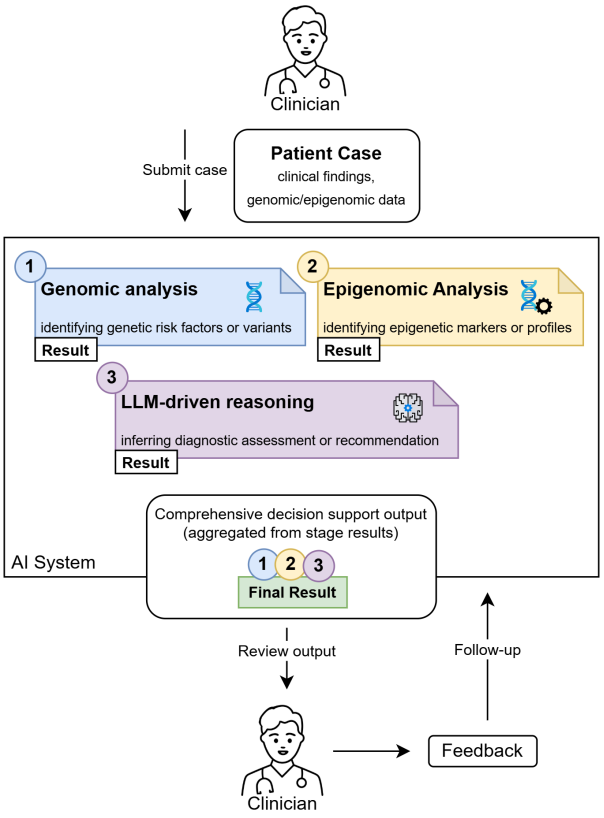


Fig. 2. Workflow of clinician interaction with an LLM-assisted diagnostic system.

scientifically grounded example, a clinician uses the LLM-supported system to interpret ambiguous genetic findings and stratify the patient's risk, leading to a personalized management plan. Figure 2 depicts the clinical workflow of this scenario, showing how the clinician interacts with the LLM system and how different LLM components contribute at each step.

Patient Background: Dr. J is evaluating a 50-year-old patient experiencing rapidly progressing memory loss and cognitive decline. The patient has a strong family history of early-onset AD – several close relatives were diagnosed in their 50s. This raises concern for a familial (possibly autosomal dominant) form of AD. Standard clinical workup, including cognitive testing and MRI, indicate early Alzheimer-type changes, but Dr. J wants a genomic analysis to clarify the risk and guide proactive treatment.

Genetic Findings: The patient undergoes a dementia-focused gene panel test. The results identify two Variants of Uncertain Significance (VUS) in the *PSEN1* gene. *PSEN1* (Presenilin-1) is well-known in AD genetics; certain mutations in *PSEN1* are causative for early-onset AD [6]. However, these particular variants are rare and not well-characterized in databases, leaving their pathogenicity unclear. Traditional resources (ClinVar, literature) have scant information on these

VUS, providing no definitive conclusions. Dr. J faces a dilemma: **how to interpret the risk associated with these variants and decide on early interventions without clear guidance from conventional sources?**.

LLM-Assisted Hypothesis Generation: Dr. J inputs the patient's genomic data and scenario into the GeneGPT pipeline, asking it to assess whether either *PSEN1* variant is likely pathogenic and what mechanisms might be involved. The genomics generative LLM quickly gathers what is known: it scans biomedical literature and databases for any mention of these specific variants or similar ones. It notes that *PSEN1* mutations generally can increase amyloid-β production in the brain (a key factor in AD pathology), referencing known research. Based on analogies to known pathogenic mutations, the LLM hypothesizes that one of the VUS (Variant A) might disrupt a critical site in the *Presenilin-1* protein, potentially leading to abnormal cleavage of amyloid precursor protein. It suggests that this variant could thus accelerate amyloid plaque formation, contributing to early cognitive decline. The other variant (Variant B), lacking such features, is hypothesized to be benign or of lower impact. The LLM provides supporting context, e.g. citing [7] "Cummings et al. (2021) noted *PSEN1* variants significantly influence AD onset" to give Dr. J biological plausible reasoning. It also proposes a possible therapeutic angle: if Variant A is indeed pathogenic, therapies targeting amyloid production or enhancing its clearance might be especially relevant for this patient.

LLM-Assisted Discriminative Validation: Next, to validate GeneGPT's assumptions, the Dr. J turns to iLLMAC to evaluate the functional impact of the *PSEN1* VUS variants refining hypotheses by discriminative analysis of phenotypic correlations [8]. Feeding in the patient's *PSEN1* sequence and known pathogenicity features, the model evaluates each variant. It classifies Variant A as "Likely Pathogenic" with a high confidence (e.g. 97% confidence). The model's explanation highlights that Variant A occurs in a functionally critical domain of the *PSEN1* protein and those multiple predictive features (e.g., evolutionary conservation, protein structure modeling) indicate a deleterious effect. It notes that such an alteration could interfere with normal amyloid processing, aligning with the generative model's hypothesis. In contrast, Variant B is classified as "Uncertain/likely benign", with the model finding insufficient evidence of harm. The discriminative LLM essentially confirms the generative LLM's suspicion about Variant A, lending credence to the idea that this variant is driving the patient's early symptoms. Dr. J now has an AI-supported classification of the previously unknown variant.

Epigenetic and Multi-omics Insights: To enrich the analysis, Dr. J also provides the patient's epigenetic data to the system (from an experimental blood test assessing DNA methylation patterns and histone marks associated with AD). EpiGePT analyzes this data and identifies an interesting pattern: the patient shows high DNA methylation at the *PSEN1* promoter region (which could downregulate *PSEN1* expression) and reduced H3K27ac histone acetylation in genes related to synaptic function. According to recent studies in AD epigenetics, such markers correlate with faster disease progression (the generative LLM even recalls a publication analogous to [9] noting these correlations). The EpiGePT produces a risk score based on the methylation/acetylation profile; the patient's profile yields a high-risk score (with a predictive AUC 0.9 in internal validation) for rapid cognitive decline. This suggests the patient's epigenetic state is aligned with aggressive disease, which is crucial information for management. Moreover, the system hypothesizes that the hypermethylation of *PSEN1* might be a compensatory response to the overactive mutant variant (an insight that connects the genomic and epigenomic findings).

Integrated Reasoning and Outcome: This integrated analysis enabled precise reclassification: one *PSEN1* variant (A) was elevated to "Likely Pathogenic," given its epigenetic impact on gene regulation, while the other variant (B) remained uncertain, warranting ongoing monitoring. The presence of high-risk epigenetic markers further stratifies the patient as high risk for rapid progression. With this knowledge, Dr. J can formulate a proactive plan. The system's output includes suggested therapeutic strategies: for example, it highlights clinical trials of γ-secretase inhibitors or anti-amyloid antibodies that specifically target the pathway affected by *PSEN1* mutations (citing ongoing trial data). It also suggests considering epigenetic therapy research, such as DNA methylation inhibitors, as a futuristic approach to modulate gene expression. In essence, the AI has not only clarified the genetic risk but also pointed to personalized interventions.

Clinician Verification and Action: Dr. J reviews the LLM-generated report. She cross-verifies key points: the classification of Variant A as Likely Pathogenic is consistent with what little is published in AD variant databases (the system helpfully references an entry in ClinVar that, while not definitive, noted this variant in one AD case). The proposed link between epigenetic markers and fast progression is intriguing; Dr. J knows that large studies (like ADNI cohorts) have reported similar findings, which aligns with the model's suggestion. Satisfied with these insights, Dr. J decides to treat the patient as high-risk AD. She enrolls the patient in a prevention trial for early-stage patients, steps up monitoring, and advises lifestyle modifications known to slow cognitive decline. She also communicates to the patient and family that advanced AI analysis has provided additional confidence in the genetic findings, which helps in their understanding and future planning. The outcome is a more informed clinical decision, achieved much faster than a manual literature review would allow.

This case study exemplifies the practical workflow of an LLM-supported neurology decision system. The illustrated Figure 2 shows the timeline of interactions: from the moment genetic data is obtained, through LLM hypothesis generation and validation, to the point where the clinician receives a comprehensive report and makes an intervention decision. The result for the patient is an earlier and more tailored intervention plan than would otherwise be possible. The scenario underscores how LLMs can turn heterogeneous data (genomic +

epigenomic + clinical) into actionable knowledge, even in the face of rare variants and uncertain data.

IV. Conclusion

The above results highlight the transformative potential of LLM-enabled decision support in neurology. By combining multiple AI models with specialized expertise, the system provided a holistic analysis that mirrors a team of specialists consulting on the case. The generative model contributed creative insight – proposing that the obscure *PSEN1* variant is pathogenic and linking it to known AD mechanisms – something a single human or algorithm might not do so readily when faced with limited prior information. The discriminative model added rigor, confirming which hypothesis held up against data. This synergy yielded a confident outcome (variant reclassification and risk stratification) that directly informed clinical action. Such a pipeline can accelerate diagnostics (reducing the time from data to decision) and possibly improve accuracy by cross-validation between models. It demonstrates that LLMs, beyond chatbots, can serve as intelligent agents in clinical workflows, ingesting complex inputs and outputting reasoned conclusions. Importantly, the LLM outputs in the scenario were accompanied by explanations (e.g., why the variant is likely pathogenic), which is crucial for clinician acceptance. This aligns with the need for AI in healthcare to be not only accurate but interpretable and aligned with clinicians' reasoning.

References

[1] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[2] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "Biogpt: generative pre-trained transformer for biomedical text generation and mining," *Briefings in bioinformatics*, vol. 23, no. 6, p. bbac409, 2022.

[3] Q. Jin, Y. Yang, Q. Chen, and Z. Lu, "Genegpt: Augmenting large language models with domain tools for improved access to biomedical information," *Bioinformatics*, vol. 40, no. 2, p. btae075, 2024.

[4] Z. Gao, Q. Liu, W. Zeng, R. Jiang, and W. H. Wong, "EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics," *Genome Biology*, vol. 25, no. 1, p. 310, Dec. 2024. [Online]. Available: https://doi.org/10.1186/s13059-024-03449-7

[5] J. Liu, H. Shen, K. Chen, and X. Li, "Large language model produces high accurate diagnosis of cancer from end-motif profiles of cell-free dna," *Briefings in Bioinformatics*, vol. 25, no. 5, p. bbae430, 09 2024. [Online]. Available: https://doi.org/10.1093/bib/bbae430

[6] R. J. Kelleher III and J. Shen, "Presenilin-1 mutations and alzheimer's disease," *Proceedings of the National Academy of Sciences*, vol. 114, no. 4, pp. 629–631, 2017.

[7] J. Cummings, G. Lee, K. Zhong, J. Fonseca, and K. Taghva, "Alzheimer's disease drug development pipeline: 2021," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 7, no. 1, p. e12179, 2021.

[8] H. Nan, M. Chu, D. Jiang, W. Liang, Y. Li, Y. Wu, Z. Wang, and L. Wu, "Identification and characterization of variants in psen1, psen2, and app genes in chinese patients with early-onset alzheimer's disease," *Alzheimer's Research & Therapy*, vol. 17, no. 1, pp. 1–18, 2025.

[9] L. M. De Plano, A. Saitta, S. Oddo, and A. Caccamo, "Epigenetic changes in alzheimer's disease: Dna methylation and histone modification," *Cells*, vol. 13, no. 8, p. 719, 2024.

Mapping the Landscape of Anonymous CTI Sharing

Asem Mousa
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University
Skopje, Republic of North Macedonia
https://orcid.org/0009-0000-9023-8925

Petre Lameski
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University
Skopje, Republic of North Macedonia
https://orcid.org/0000-0002-5336-1796

***Abstract*—This research systematically examines the landscape of Cyber Threat Intelligence (CTI) sharing, addressing privacy risks and proposing decentralized solutions. The study identifies key challenges in CTI dissemination, including data sensitivity, interoperability, and legal constraints, while evaluating current approaches such as blockchain-based models, machine learning (ML), and anonymization techniques. Through a structured literature review of relevant studies, the paper highlights the need for secure, privacy-preserving CTI sharing frameworks that balance anonymity with actionable intelligence. Findings reveal that blockchain and AI-driven automation enhance trust and efficiency in CTI exchange, but gaps remain in scalability and regulatory compliance. The study concludes by advocating for adaptive anonymization mechanisms, decentralized collaboration models, and standardized privacy-preserving protocols to improve CTI sharing in cybersecurity defense.**

***Index Terms*—Cyber Threat Intelligence, CTI, Anonymous sharing, Privacy, Cybersecurity**

I. Introduction

The number of reported cyberattacks, such as ransomware, phishing, and social engineering, has increased significantly in recent years, drawing the attention of both the public and commercial sectors [3]. Cyber Threat Intelligence (CTI) has emerged as a key player in modern cybersecurity defense mechanisms, allowing organizations to predict, detect, and respond to threats more effectively. Cyber Threat Intelligence (CTI) involves analyzing data to generate actionable insights about existing or emerging cyber threats targeting an organization. By understanding attackers' motives and capabilities, CTI enables organizations to proactively prevent or mitigate cyberattacks, shifting from a reactive to a proactive security approach. This approach improves decision making, strengthens defense mechanisms, and reduces potential financial and reputational damage [7].

The primary goal of CTI is to transform raw data into meaningful intelligence that can inform decision making and improve the security posture of an organization. By analyzing indicators of compromise (IoCs), threat actors' tactics, techniques, and procedures (TTPs), and broader trends in the cyber threat space, CTI empowers organizations to identify vulnerabilities, prioritize remediation process, and respond to incidents with greater speed and precision [4]. For example, CTI can reveal patterns in phishing campaigns, malware distribution, or ransomware attacks, allowing organizations to implement targeted defenses and reduce their risk exposure [16]. This intelligence-driven approach is particularly critical in an era where cyber threats not only are increasing in frequency but are also becoming more sophisticated and difficult to detect.

The value of CTI goes beyond mere threat detection; it also plays a crucial role in strategic planning and risk management. By understanding the motivations and capabilities of threat actors, organizations can better allocate resources, develop robust incident response plans, and strengthen their overall security infrastructure [7]. For example, CTI can help organizations identify which assets are most likely to be targeted, enabling them to implement proper security measures and reduce the likelihood of a successful attack. Furthermore, CTI fosters collaboration within the cybersecurity community, as organizations share threat data and insights to collectively combat cybercrime [16].

Despite its many benefits, the implementation of CTI is not without challenges. The big amount of data generated by modern IT environments can overwhelm security teams, making it difficult to distinguish between relevant threats and false positives. Furthermore, the dynamic nature of the cyber threat landscape requires organizations to continuously update their intelligence-gathering and analysis capabilities to stay ahead of adversaries [9]. To address these challenges, many organizations are turning to automated tools and platforms that leverage artificial intelligence (AI) and machine learning (ML) to streamline the CTI process and improve its accuracy [4].

Cyber Threat Intelligence reports contain many Indicator of Compromise (IoC) data, which might reveal sensitive information about the organization's infrastructure, such as IP addresses, file hashes and network domains. Moreover, it could violate the privacy of users by revealing usernames, emails, and IP addresses [8]. The research objective is to address the following questions:

- What is the privacy risk of extracting Personally Identifiable Information (PII) and sensitive information from threat intelligence reports?
- How can a decentralized collaborative model improve privacy and security in the sharing of cyber threat intelligence (CTI) using anonymization mechanisms?

The rest of the paper is organized as follows: Section II discusses the methodology and criteria for paper selection. Section III presents a summary of the literature review. Section IV discusses the findings and results. Finally, section

V concludes the study, addressing research gaps and future directions.

II. Method

A scoping review aims to identify and evaluate key sources related to a specific research topic. This method is a thorough and well-established procedure that is essential for conducting any research study. The current review follows the guidelines set forth by [10] and draws upon previous systematic reviews.

The reviewed studies are categorized based on their research area. Most of the research we reviewed included surveys and challenges related to CTI sharing, in addition to conducted studies on different approaches of implementing CTI.

A. Selection criteria

When choosing the papers for our literature review I, we follow the scoping review guidelines explained in [10], which includes defining research questions and objectives, specifying criteria, collecting evidence, and analyzing results. We ensure that only relevant articles will be included in the scoping review. Each critically assessed study should cover at least one of the following topics:

- CTI challenges
- Safe Threat Intelligence sharing
- CTI data collection methods
- Indicators of Compromise-feeds

Multiple papers discussed CTI data sharing in terms of different mechanisms and methods. While others examined current frameworks and sharing standards, this included surveys and reviews of available platforms.

B. Data sources and search strategies

The studies included in this systematic review were collected from various online databases, including PubMed, IEEE Xplore, Science Direct, and Google Scholar. The search keywords employed were (”Cyber Threat Intelligence sharing” OR ”Cyber Threat Intelligence” OR ”Data sharing privacy”). The articles were extracted in March 2025. The search results yielded 1130 studies using these keywords. Among these, we chose the selected ones in accordance with our criteria, aligning with the primary objective of this review. In summary, the total number of included studies is 9. This comprise research articles from journals, books, and conference papers. The articles were meticulously analyzed to ensure the reliability of the findings. Discrepancies in the analyses were resolved through discussions and additional evaluations of the contested research.

III. Results and Discussion

This section presents the papers in Literature Review Table I, which were selected according to the selection criteria mentioned in the II-A. Each of the selected papers is calssified according to its category 1, as most of the literature included surveys about state-of-art Cyber Threat Intelligence sharing and its challenges, while the rest involved actual implementation of CTI framework, models, or platforms.

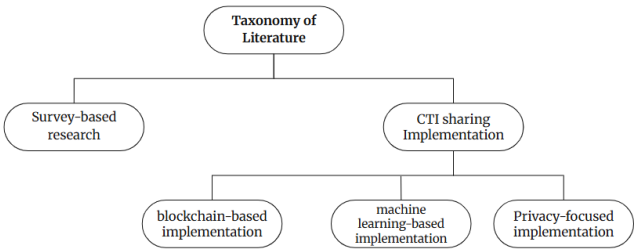


Fig. 1: Taxonomy of Literature

A. Threat Intelligence Sharing Surveys

- The authors in [2] explores how Cyber Threat Intelligence (CTI) is shared, focusing on current methods and where future research could go. CTI plays a key role in reducing cyber threats by providing real-time data. Today’s approaches include blockchain, incentive-driven models, machine learning, and advanced computing platforms. Moving forward, research should explore AI-driven automation, large-scale data management, adaptable identity systems, evolving security regulations, and even quantum computing. The report emphasizes the need for further studies to develop CTI sharing frameworks that are robust, scalable, and secure.
- The paper in [12] investigates Cyber Threat Intelligence (CTI)’s function in improving organizational cybersecurity resilience. It examines 52 peer-reviewed papers from 2019 to 2023, highlighting major trends, techniques, and frameworks in CTI. The review organizes research into detection models and knowledge-sharing activities. A three-layered CTI architecture is suggested, consisting of a knowledge base, detection models, and visualization dashboards. Issues such as data overload and legal limits are considered. Future directions include customized CTI installations and more stakeholder involvement.
- The study in [1] underlines the need of sharing Cyber Threat Intelligence (CTI) to improve organizational cybersecurity. Traditional signature-based defense systems are ineffective against evolving threats, necessitating real-time CTI to reduce detection gaps. CTI sources include open-source feeds, dark web data, and commercial suppliers. TIS platforms provide uniform exchanges, but trust challenges, legal and privacy concerns, data overload, and high implementation costs impede wider adoption. The study advocates for more organizational involvement in TIS to create a proactive, collaborative cybersecurity environment.
- The research [15] investigates the significance of sharing Cyber Threat Intelligence (CTI) in the fight against cyber attacks. It emphasizes the importance of standardized automated procedures for successful CTI sharing, as well as technical and non-technical challenges. The report also identifies six key characteristics of threat intelligence: trustworthiness, reputation evaluation, contextual relevance, adequate anonymization, timely distribution,

TABLE I: Summary of Literature Review on Anonymous and General CTI Sharing

Reference	Year	Method	Contribution	Authors
Current approaches and future directions for Cyber Threat Intelligence sharing: A survey [2]	2024	Literature review, Survey	Comprehensive survey on CTI sharing, analyzing architectures, challenges, solutions, and future directions.	Poopak Alaeifar, Shantanu Pal, Zahra Jadidi, Mukhtar Hussain, Ernest Foo
Standards-based Cyber Threat Intelligence sharing using private Blockchains [11]	2023	Conceptual framework, experimental analysis	Develops a decentralized CTI platform via Hyperledger Fabric using STIX/TAXII to enhance security, auditability, and trust.	Provatas, Kimonas; Tzannetos, Ioannis; Vescoukis, Vassilios
Blockchain-Based Cyber Threat Intelligence Sharing Using Proof-of-Quality Consensus [5]	2023	Practical validation, comparative analysis	Introduces PoQ consensus mechanism with validator selection based on reputation and quality to enhance trust.	Chatziamanetoglou, Dimitrios; Rantos, Konstantinos
A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience [12]	2023	Systematic Literature Review	Proposes a CTI framework for detection and risk visualization to support organizational cybersecurity hardening.	Saeed, Saqib; Suayyid, Sarah A.; Al-Ghamdi, Manal S.; Al-Muhaisen, Hayfa; Almuhaideb, Abdullah M.
Robust Botnet DGA Detection: Blending XAI and OSINT for Cyber Threat Intelligence Sharing [13]	2022	Quantitative analysis, data collection	Proposes botnet DGA detection using statistical features; blends XAI and OSINT for explainable, bias-resistant intelligence.	Suryotrisongko, Hatma; Musashi, Yasuo; Tsuneda, Akio; Sugitani, Kenichi
A Theoretical Review on the Importance of Threat Intelligence Sharing & The Challenges Intricated [1]	2021	Descriptive, literature review	Highlights existing platforms and the need for automation and standardization in CTI sharing.	Sandhya Sukhabogia; Dr. M. Anushab
An Overview of Cyber Threat Intelligence Platform and Role of AI and Machine Learning [6]	2020	Literature review, data analysis	Proposes AI/ML-enhanced CTI framework for improved intelligence extraction.	Abir Dutta; Shri Kant
Cyber Threat Intelligence Sharing: Survey and Research Directions [15]	2019	Literature survey	Identifies challenges, trends, and the need for automation in CTI sharing to enhance cyber defense.	Wagner, Thomas D.; Mahbub, Khaled; Palomar, Esther; Abdallah, Ali E.
Towards an Anonymity-Supported Platform for Shared Cyber Threat Intelligence [14]	2018	Prototype development, comparative analysis	Develops anonymity-preserving CTI platform using regex-based PII masking and TOR integration.	Wagner, Thomas; Palomar, Esther; Mahbub, Khaled; Abdallah, Ali

and system interoperability. The authors propose future research into trust-building systems, intelligent relevance filtering techniques, and legally permissible sharing processes.

B. Threat Intelligence Sharing implementation

There are many apporoaches when it comes to implement CTI sharing such as using blockchain technologies and machine learning.

1) blockchain-based implementation:

- The authors in [11] presents a decentralized approach to sharing Cyber Threat Intelligence (CTI) using blockchain technology. The system, built on Hyperledger Fabric, offers stronger confidentiality, data integrity, availability, and accountability compared to traditional centralized models. It integrates REST API gateways, smart contracts, and CouchDB for efficient STIX data storage. The team plans to enhance the system with on-chain STIX validation and multi-factor authentication for added security. This approach could significantly improve real-world operations by preserving compatibility with industry standards.
- The paper [5] presents a blockchain-based system for sharing and evaluating cyber threat intelligence (CTI) using a Proof-of-Quality consensus algorithm. The model, Awareness Architecture Based on Blockchain CTI Convergence (ABC)², consists of three roles: CTI Feed Sources, Validators, and Consumers. Validators are chosen based on reputation metrics stored on the blockchain. The PoQ consensus algorithm uses a voting mechanism to rate CTI feeds against predefined quality parameters. The system stores CTI feeds, evaluations, and validator performance in an immutable ledger. The model's re-

silience against malicious validators is demonstrated in a theoretical simulation. Future work includes a proof-of-concept implementation and integration with real-time risk management frameworks.

2) machine learning-based implementation:

- Proposal [13] introduced an approach to identifying botnet domain generation algorithm (DGA) traffic that combines explainable artificial intelligence (XAI) with open-source intelligence (OSINT). The scientists analyzed 12 years of DNS query information to create a machine learning model that is 96.3% accurate across 55 DGA families. They suggest the notion of "computable CTI," in which AI/ML models and danger indicators are shared to increase transparency. By combining XAI approaches with OSINT validation, the strategy offers security analysts with unambiguous, interpretable forecasts and second-hand threat assessments. The study demonstrates how integrating XAI and OSINT may improve security decision-making and facilitate collaborative threat intelligence sharing.
- The study in [6] introduced the use of artificial intelligence (AI) and machine learning in CTI systems to strengthen cybersecurity defenses against emerging threats. A supervised ML model with a Naïve Bayes classifier can extract actionable threat intelligence from both structured and unstructured data. Natural language processing methods are used in the model to handle text-based threat data. The paper finds shortcomings in existing CTI methods and proposes future improvements. The framework intends to expedite threat intelligence integration into current security systems, allowing for faster detection and mitigation of cyber threats while minimizing reliance on human analysis.

Other implementations involved a proposal [14] that presents a CTI sharing platform for exchanging cyber threat intelligence (CTI) that includes built-in anonymity capabilities to meet privacy and security concerns. The Malware Information Sharing Platform (MISP) and the Traffic Light Protocol (TLP) are used to disguise personally identifiable information (PII) with adjustable regular expressions. White-labeled data is anonymized to the greatest extent possible before being shared with the public. To avoid traceability, the system directs connections over the TOR network using a Whonix gateway. The prototype is highly accurate in anonymizing multiple sorts of PII, with just three current platforms providing partial anonymity characteristics. The authors suggest that this method promotes safe CTI sharing by automating privacy preservation while retaining data value.

IV. Conclusion

This research highlights the need for anonymous CTI sharing in modern cybersecurity, emphasizing the necessity to mitigate privacy risks while maintaining intelligence utility. The review reveals that blockchain-based frameworks (e.g., Hyperledger Fabric, Proof-of-Quality consensus) and AI/ML-driven detection models (e.g., botnet DGA analysis, NLP-based threat extraction) significantly enhance CTI sharing by improving trust, automation, and data integrity. However, challenges such as PII exposure, legal constraints, and interoperability gaps persist, limiting widespread adoption. Future research should focus on scalable anonymization techniques, decentralized trust models, and standardized sharing mechanisms to foster secure and efficient CTI collaboration. By addressing these gaps, organizations can transition toward proactive, privacy-aware threat intelligence ecosystems, strengthening collective cyber defense against evolving threats.

V. Acknowledgment

The work presented in this paper was partially funded by the Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering. And was supported partially by the European Union in the framework of ERASMUS MUNDUS, Project CyberMACS (Project #101082683) (https://cybermacs.eu).

References

[1] S. S. et al. "A Theoretical Review on the Importance of Threat Intelligence Sharing & The Challenges Intricated". In: *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.3 (Apr. 2021), pp. 3950–3956.

[2] Poopak Alaeifar et al. "Current approaches and future directions for Cyber Threat Intelligence sharing: A survey". In: *Journal of Information Security and Applications* 83 (2024), p. 103786. ISSN: 2214-2126. DOI: https://doi.org/10.1016/j.jisa.2024.103786. URL: https://www.sciencedirect.com/science/article/pii/S2214212624000899.

[3] Fatimah Aldauiji, Omar Batarfi, and Manal Bayousef. "Utilizing Cyber Threat Hunting Techniques to Find Ransomware Attacks: A Survey of the State of the Art". In: *IEEE Access* 10 (2022), pp. 61695–61706. DOI: 10.1109/ACCESS.2022.3181278.

[4] Kurt Baker. *Cyber Threat Intelligence Explained*. 2025. URL: https://www.crowdstrike.com/en-us/cybersecurity-101/threat-intelligence/.

[5] Dimitrios Chatziamanetoglou and Konstantinos Rantos. "Blockchain-Based Cyber Threat Intelligence Sharing Using Proof-of-Quality Consensus". In: *Security and Communication Networks* 2023.1 (2023), p. 3303122.

[6] Abir Dutta and Shri Kant. "An Overview of Cyber Threat Intelligence Platform and Role of Artificial Intelligence and Machine Learning". In: *Information Systems Security*. Springer Cham, Dec. 2020, pp. 81–86. ISBN: 978-3-030-65609-6. DOI: 10.1007/978-3-030-65610-2_5.

[7] eccouncil. *What is Threat Intelligence in Cybersecurity?* 2024. URL: https://www.eccouncil.org/cybersecurity-exchange/threat-intelligence/what-is-cyber-threat-intelligence/.

[8] Gina Fisk et al. “Privacy Principles for Sharing Cyber Security Data”. In: May 2015, pp. 193–197. DOI: 10.1109/SPW.2015.23.

[9] kaspersky. *What is threat intelligence? Definition and explanation*. 2025. URL: https://www.kaspersky.com/resource-center/definitions/threat-intelligence (visited on 03/25/2025).

[10] Danielle Pollock et al. ““How-to”: scoping review?” In: *Journal of Clinical Epidemiology* 176 (2024), p. 111572. ISSN: 0895-4356. DOI: https://doi.org/10.1016/j.jclinepi.2024.111572. URL: https://www.sciencedirect.com/science/article/pii/S0895435624003287.

[11] Kimonas Provatas, Ioannis Tzannetos, and Vassilios Vescoukis. “Standards-Based Cyber Threat Intelligence Sharing Using Private Blockchains”. In: *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. 2023, pp. 649–656. DOI: 10.15439/2023F6880.

[12] Saqib Saeed et al. “A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience”. In: *Sensors* 23.16 (2023). ISSN: 1424-8220. DOI: 10.3390/s23167273. URL: https://www.mdpi.com/1424-8220/23/16/7273.

[13] Hatma Suryotrisongko et al. “Robust Botnet DGA Detection: Blending XAI and OSINT for Cyber Threat Intelligence Sharing”. In: *IEEE Access* 10 (2022), pp. 34613–34624. DOI: 10.1109/ACCESS.2022.3162588.

[14] Thomas Wagner et al. “Towards an Anonymity Supported Platform for Shared Cyber Threat Intelligence”. In: *Risks and Security of Internet and Systems*. Springer International Publishing, Jan. 2018, pp. 175–183. ISBN: 978-3-319-76686-7. DOI: 10.1007/978-3-319-76687-4_12.

[15] Thomas D. Wagner et al. “Cyber threat intelligence sharing: Survey and research directions”. In: *Computers Security* 87 (2019), p. 101589. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2019.101589. URL: https://www.sciencedirect.com/science/article/pii/S016740481830467X.

[16] Hemant Warier. *The Crucial Role of Cyber Threat Intelligence in Protecting Your Organization*. 2025. URL: https://www.cloudsek.com/knowledge-base/the-crucial-role-of-cyber-threat-intelligence-in-protecting-your-organization#:~:text=The%20use%20of%20threat%20intelligence,their%20assets%20from%20potential%20breaches. (visited on 02/13/2025).

Hate Speech Detection on Social Media Using Transformer-Based Models: A BERT Approach

Mihail Trajkovikj
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, N. Macedonia
mihail.trajkovikj@students.finki.ukim.mk

Marko Lazarevski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, N. Macedonia
marko.lazarevski@students.finki.ukim.mk

Anabela Bochvarova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, N. Macedonia
anabela.bochvarova@students.finki.ukim.mk

Evgenij Anushev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, N. Macedonia
evgenij.anushev@students.finki.ukim.mk

Ilija Trajkovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, N. Macedonia
ilija.trajkovski@students.finki.ukim.mk

Georgina Mirceva
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, N. Macedonia
georgina.mirceva@finki.ukim.mk

***Abstract*—In this paper, we take a closer look at the challenge of detecting hate speech on social media. The rise in online hate speech and its harmful effects on society is what drove us to explore this issue. We reviewed several important studies in the field, focusing on the methods they used and their performance. In this paper, we use a BERT transformer model for solving the task at hand. We use a pre-trained BERT model and fine tune the model to make accurate hate speech detection. We present how data pre-processing is made, how the model is trained and also, we present the performances of the model. Although hate speech detection models achieve good results, there's still a lot of work to be done.**

Keywords—hate speech detection, deep learning, BERT model

I. Introduction

Social media has completely changed the way we communicate with each other. It made it easy to connect and share ideas with people from all over the world, but it also opened the door to a rise in hate speech. Whether it's targeting people because of their race, religion, gender, or sexual orientation, hate speech can lead to violence, discrimination, and create a toxic space online. Because of how serious the consequences can be, there's a huge need for detection tools that can spot and stop hate speech in real time.

Over the years, researchers have tried different ways to tackle this problem. Some used old-school rule-based systems that use a list with some predefined keywords, while others moved toward more advanced machine learning and deep learning models that can recognize patterns in language. Each method has its advantages and disadvantages, and their success often depends on the specific context and data they're dealing with.

Plenty of studies have already looked into hate speech detection, each using different techniques and data sources. This problem has been addressed with various machine learning (ML) and deep learning (DL) techniques. The former approaches apply feature extraction, and then use some ML technique for making decisions [1]. Approaches based on bag of words and n-grams [2] are used, as well lexicon based approaches that use a predefined list of related terms [3].

Besides the traditional machine learning techniques, also DL techniques are used for solving this task [4]. For example, word embedding techniques can be used for feature extraction [5], [6], and then recurrent neural networks (RNNs) [7] and/or convolutional neural networks (CNNs) [8] could be applied. In this way, the researchers managed to get better results. These findings highlight the need for more sophisticated techniques if we want more accurate models.

Lately, advances in natural language processing (NLP) have brought us powerful tools like transformer-based models, which have also been applied for hate speech detection [9]. For example, the Bidirectional Encoder Representations from Transformers (BERT) model [10] can be used, which have shown good performance for many NLP tasks.

Even with all the progress so far, there are still big hurdles. Language can be tricky—context, cultural differences, and the rapid evolution of slang make it hard to keep up. Models can struggle to tell if something is offensive or actually hate speech. Solving these issues is key to building more reliable systems.

Detecting hate speech comes with ethical challenges too. Balancing freedom of speech with the need to protect people from harmful language isn't easy. Researchers have to be careful that their models don't accidentally silence legitimate conversations while targeting harmful content. Ethical considerations like these are super important when developing and using detection tools.

In this paper, we use a pre-trained BERT model and fine tune for solving the task for hate speech detection. The rest of this paper is organized as follows. In section 2, we present the approach that is used. We give details about the dataset that is used, the pre-processing steps that are applied over the data, as well as how the model is trained. Section 3 give evidence about the experimental results that are obtained, while section 4 gives the final conclusions, as well as directions for further research.

II. Our Approach

A. Dataset

In this research, we used a labeled dataset [11], [12] of tweets that were classified into three categories: hate speech, offensive language, and neither. The dataset includes 24783 tweets, which are distributed among the three classes. We want to note that the dataset is not balanced. Namely, 11.6% of the samples are labelled as hate speech, 76% are labelled as offensive language and the remainder are labeled as neither (non-offensive). The dataset is divided into training and validation set. The training set contains 80% of the samples, while the validation set contains the remaining 20% of the samples.

B. Data Pre-processing

Before training the model, we took several important steps to clean and prepare the data. These pre-processing steps help to ensure that the model could better understand the text and focus on the important features. The following pre-processing steps are applied:

- **Lowercasing**: Converting all text to lowercase in order to ensure uniformity;
- **Removing Special Characters**: Eliminating URLs, mentions, and special characters that do not contribute to the sentiment;
- **Tokenization**: Splitting the text into individual words or tokens;
- **Stop Word Removal**: Removing common words that do not add significant meaning to the text, such as "and," "the," "is." etc.;
- **Lemmatization**: Reducing words to their base or root form to improve model performance (ex. convert “running” to “run”).

C. BERT Model

The dataset was split into training and test set, where 80% served as training data, and the remaining samples were used for validation of the model’s performance. We used a BERT model [10] because of its proven success in NLP tasks. The model is trained with the following steps:

- **Model Initialization**: We loaded a pre-trained BERT model and tokenizer;
- **Data Encoding**: We converted the text into input IDs and attention masks that BERT can work with;
- **Training**: We used the AdamW optimizer with a learning rate of $5e^{-5}$ and a batch size of 16. We trained the model for 4 epochs.

After training, we tested the model on the validation set to see how well it performed.

III. Evaluation

We made several experiments. First, we trained standard model, while in the second experiment we tried to improve the prediction power of the model by giving different weights in the decision about the classes in order to work in “balanced mode” as if the data are balanced. The experimental results from these experiments are shown in Table 1 and Table 2, respectively.

TABLE I. Experimental Results Obtained with the First Experiment (Without Class Weighting)

Class	Evaluation Measure		
	Precision	***Recall***	***F1-measure***
hate speech	0.48	0.47	0.48
offensive language	0.95	0.93	0.94
neither	0.82	0.87	0.85
Other Evaluation Measures			
macro avg	0.75	0.76	0.75
weighted avg	0.90	0.90	0.90

TABLE II. Experimental Results Obtained with the Second Experiment (With Class Weighting)

Class	Evaluation Measure		
	Precision	***Recall***	***F1-measure***
hate speech	0.32	0.58	0.41
offensive language	0.96	0.85	0.90
neither	0.73	0.92	0.81
Other Evaluation Measures			
macro avg	0.67	0.78	0.71
weighted avg	0.89	0.85	0.86

From the results, we can notice that the precision is generally lower in the second experiment. Namely, more samples are classified as hate speech or neither that do not belong to the corresponding class, which also result in decrease in the recall for the offensive speech class since these samples are classified in the hate speech and neither classes. On the other side, the recall for the hate speech and neither classes is higher in the second experiment, leading to lower precision for these classes, but on the other hand, for the offensive language class we have lower recall with a slight increase on the precision.

The macro average measures shows that in general, the recall is better in the second experiment, while the precision is better in the first experiment. If we look at the weighted average measures, where the dominant class would get higher weight, there is slight decrease in the precision and more significant decrease in the recall in the second experiment, proving the same that the dominant class that would get higher weight is predicted worse in the second experiment.

Finally, if we look at the values for the F1-measure, for all classes, with the first experiment higher results are obtained. Also the macro average and weighted average F1-measure is higher in the first experiment.

However, considering that the task at hand is hate speech detection, it is more important to not dismiss a hate speech sample and to present it to the end-user. This means that the false negatives for this class should be as low as possible, meaning that the recall for the hate speech class is most important. Also, the recall for the offensive language class would be important if we want to successfully detect such samples in order not to show the text to the end-user.

TABLE III. EXPERIMENTAL RESULTS OBTAINED WITH THE THIRD EXPERIMENT (WITH DATA BALANCING)

Class	Evaluation Measure		
	Precision	***Recall***	***F1-measure***
hate speech	0.98	1.00	0.99
offensive language	0.97	0.96	0.96
neither	0.91	0.87	0.89
Other Evaluation Measures			
macro avg	0.95	0.94	0.95
weighted avg	0.97	0.97	0.97

From the results, it is evident that for the hate speech class there is an improvement in the recall using class weighting in the model, but it also resulted in lower recall for the dominant class (offensive language class). Also for the neither class, the recall is better with the second experiment.

The macro average recall is slightly better when using class weighting, but it also significantly decreases the macro average precision. Therefore, further experiments can be made, where also balancing of the data can be made in order to see if the recall for the hate speech and offensive language classes could be improved.

Therefore, we also made the third experiment, where balancing of the data is made, particularly for hate speech and offensive language classes. The results are shown in Table 3. From the results, we can see that for all classes there is an improvement of the results obtained for all measures. Exclusion from this is the recall for the neither class, meaning that more samples from this class are classified in the hate speech and offensive language classes. However, we want to note that neither class is represented with a smaller number of representatives in the test set in this experiment, which can be seen from the confusion matrix given in Fig. 1. From Fig. 1, we can see that the model successfully predicts the hate speech and offensive language classes, having high precision and recall for both classes, but for neither class, some samples are misclassified as offensive language, and also there are larger number of samples from the offensive language class that are classified in the neither class. This is because the balancing is made in order to balance hate speech and offensive language classes, so still there is some bias towards them that could be further improved. From the micro and macro averaging measures, we can conclude that the model obtained with data balancing is significantly better, leading to higher precision and recall.

In Fig. 2, we present the training and validation loss through the epochs obtained with the third experiment. It is evident that the model is not over-fitted, and the validation loss remains approximately the same after 2 epochs.

The results indicate that while the BERT-based model performs well in detecting hate speech, there are still areas for improvement. The misclassification of offensive language as hate speech suggests that the model may benefit from additional training data or more nuanced feature extraction techniques. Adding context like information about the user or the conversation thread—might help the model understand the messages even better.

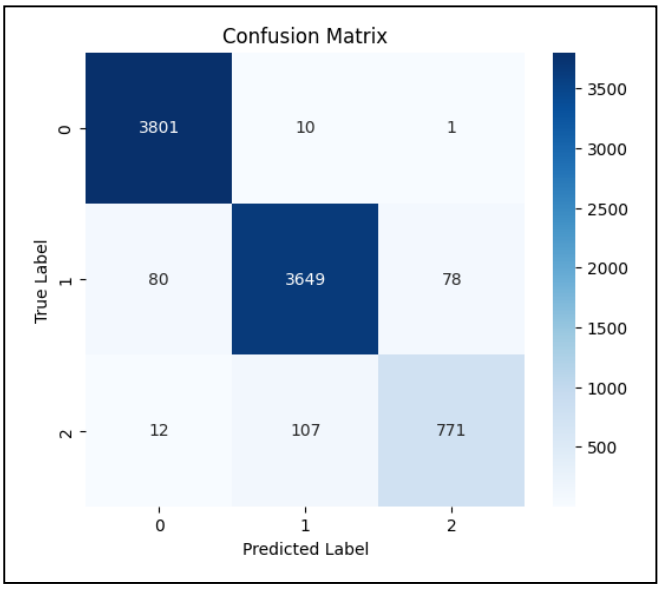


Fig. 1. Confusion matrix from the third experiment (with data balancing).

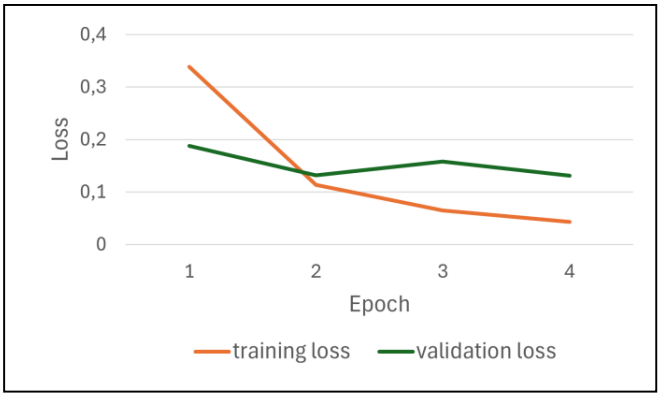


Fig. 2. Training and validation loss in the third experiment (with data balancing).

IV. CONCLUSION

In conclusion, our findings indicate that BERT-based models can effectively detect hate speech in social media. However, challenges remain, particularly in addressing the nuances of language and context.

Future work may explore the integration of additional contextual information and the use of ensemble methods to enhance detection accuracy. Additionally, expanding the dataset to include more diverse examples of hate speech could improve the model's robustness.

Future research should focus on several key areas to enhance hate speech detection systems. The user context could be incorporated as well as the conversation history to improve the accuracy of the model. Also, multi-language models could be developed that can detect hate speech in multiple languages, considering cultural differences. In order to fully use the benefits of such models, real-time detection is important, meaning that the systems should analyze and flag hate speech in real time, providing immediate feedback to users and moderators. In order to further improve the models,

also user feedback should be considered. Therefore, user feedback mechanisms should be provided thus creating feedback loops where users can report misclassifications, allowing the model to learn and adapt over time that could lead to more accurate models.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje, Skopje, N. Macedonia.

REFERENCES

[1] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," Fifth international workshop on natural language processing for social media, pp. 1–10, 2017.

[2] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," 25th international conference on world wide web, pp. 145–153, 2016.

[3] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 4, pp. 215–230, 2015.

[4] J. S. Malik, G. Pang, and A. v. d. Hengelm "Deep learning for hate speech detection: a comparative study," arXiv: 2202.09517, 2022.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Advances in neural information processing systems, pp. 3111–3119, 2013.

[6] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," The 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543, 2014.

[7] Z. Mossie and J. -H. Wang, "Vulnerable community identification using hate speech detection on social media," Information Processing & Management, vol. 57, no. 3, p. 102087, 2020.

[8] B. Gamback and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," First workshop on abusive language online, pp. 85–90, 2017.

[9] M. R. Awal, R. Cao, R. K. -W. Lee, and S. Mitrovic, "Angrybert: Joint learning target and emotion for hate speech detection," Advances in: 25th Pacific-Asia Conference, PAKDD 2021, Proceedings, Part I, pp. 701–713, Springer, 2021.

[10] J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, vol. 1, pp. 4171-4186, 2019.

[11] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," arXiv:1703.04009, 2017.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Hate Speech and Offensive Language Dataset", https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset, Accessed: March, 25, 2025.

Access Control Using Threshold Cryptography and Dynamic Policy Management

Aldiyar Ismailov
FINKI
UKIM
Skopje, North Macedonia
aldiyar.ismailov@students.finki.ukim.mk

***Abstract*—Traditional access control systems often rely on static authentication mechanisms like PINs or RFID cards, proving inadequate in today's dynamic threat landscape. This paper proposes a work-in-progress framework for an adaptive access control system, specifically developed for interacting with a smart doorphone scenario. The system is planned to integrate threshold cryptography for decentralized decision-making, Natural Language Processing (NLP) for intuitive user interaction, dynamic policy management based on trust scores, and anomaly detection. This approach aims to enhance security and usability compared to static systems by distributing authorization control among residents via cryptographic voting and dynamically adjusting access policies based on contextual factors and behavioral analysis. This paper outlines the motivation, research questions, proposed methodology, and initial considerations for technical feasibility and limitations.**

***Index Terms*—access control, NLP, cryptography, implementation**

I. Introduction

A. Motivation

Static access control methods struggle with the complexities of real-world scenarios and evolving security threats. A simple request for entry, for instance, involves variables like the requester's identity and purpose, the time of day, and the consensus of residents – factors that are not predefined. The motivation for this research stems from the need for more adaptive, resilient, and context-aware access control systems. Current systems often fail to balance security with usability, particularly in residential settings where interactions involve diverse individuals like residents, guests or delivery personnel. There's a clear gap in systems that can leverage collective resident input and dynamically adapt security postures based on real-time interaction analysis.

B. Contribution

This thesis proposes a proof-of-concept system that integrates threshold cryptography, dynamic policy management informed by trust scoring and anomaly detection, and NLP-driven interaction. The paper contribute by demonstrating a framework unifying decentralized cryptographic voting using threshold cryptography with dynamic, trust-based access policies. Also integration of an NLP interface to take voice commands from visitors and extract contextual information to inform access decisions. Additionally it will all be equipped with a mechanism for dynamically adjusting trust scores based on interaction history and anomaly detection, influencing access privileges.

C. Research Questions

The overall goal of this thesis is to research and develop a more adaptive and secure access control system. Key research questions include:

- How can threshold cryptographic protocols be effectively implemented for decentralized access authorization in a residential smart doorphone system?
- How can dynamic trust models be developed and integrated to adjust access policies based on user behavior, contextual information and detected anomalies?
- How can NLP be utilized to create an intuitive interaction layer and provide contextual data that dynamically influences the threshold voting parameters or trust scores?
- What are the security implications and performance trade-offs of integrating these technologies in an access control system?

II. Background

Before diving into the specifics of our proposed system, it's important to understand the key technologies and concepts it builds upon. This chapter provides the necessary background information. I will explore threshold cryptography, which allows for secure group decisions. I will look at Natural Language Processing (NLP), enabling computers to understand human language. I will discuss modern Trust Management Models that allow security systems to adapt based on behavior. Finally, I will cover Anomaly Detection techniques used to spot unusual or suspicious activities.

A. Threshold Cryptography

Imagine needing a highly secure key or the authority to perform a critical action, but you don't want to entrust it to a single person or system component. Threshold cryptography offers a solution by distributing this power. Instead of one entity holding the master key, the key is split into multiple pieces, called shares. These shares are distributed among 'n' different participants or devices.

The core idea is the "(t, n)-threshold". This means that to reconstruct the original secret or perform the critical action

like authorizing access, at least 't', the threshold number, out of the 'n' participants must collaborate and combine their shares. Any group smaller than 't' cannot succeed; they gain no useful information about the secret. This setup significantly increases security and resilience. A compromise of fewer than 't' participants does not break the system, and the system can still function even if some participants are unavailable.

B. NLP

NLP is a branch of artificial intelligence focused on enabling computers to process, understand, and interpret human language, similar to how humans do. Since humans naturally communicate using words and sentences, NLP is essential for creating intuitive interfaces between people and technology. In our project, NLP bridges the gap between a visitor speaking naturally to the doorphone and the system's logical processes.

C. Trust Management Models

In cybersecurity, "trust" isn't about emotion; it's a calculated assessment of how likely an entity like a user, a device, or another system is to behave correctly and securely based on its identity, attributes, and past behavior. Traditional security systems often operated on a simple, static model: once you successfully logged in at the network edge, you were generally trusted inside. This approach has proven vulnerable.

D. Anomaly Detection

Anomaly detection is the process of identifying items, events, or observations that do not conform to an expected pattern or other items in a dataset. In the context of security and access control, anomalies often signal problems – perhaps a system malfunction, a user making a mistake, or even an attacker attempting to breach the system. Detecting these deviations early allows the system to react, perhaps by raising an alert, blocking an action, or lowering a trust score.

III. State of the Art / Related Work

To understand where our project fits in, we need to look at existing research. This section reviews current knowledge in several related areas. We explore threshold cryptography, especially its use in voting systems. We examine how NLP is used in security, particularly for the Internet of Things (IoT). We also look at modern ways to control access, including dynamic trust models. Finally, we identify a specific gap in current research that this thesis aims to fill. Key aspects of the findings paper by paper will be summarized in the table later on.

A. Threshold Cryptography and Electronic Voting

You have already imagined needing a group decision where security is vital, and you don't want any single person holding all the power. Threshold cryptography offers tools for exactly this situation. It allows us to split a secret capability, like the power to unlock something or sign a document, among several people [1]. A certain number of people must cooperate to use the capability, but smaller groups cannot [2]. This prevents abuse and ensures the system still works even if some participants are unavailable or compromised.

Two well-known methods for splitting secrets are Shamir's Secret Sharing, which uses math based on polynomials [2], and the Asmuth-Bloom scheme, which uses the Chinese Remainder Theorem [3]. These methods allow shares of a secret to be distributed. Later, if enough shares are brought together, the original secret or capability can be reconstructed or used. The NIST has published detailed reports on how these threshold schemes work for various cryptographic tasks [1].

A major application of these ideas is in secure electronic voting or e-voting. Researchers have developed many e-voting systems that use threshold cryptography to protect vote secrecy and ensure accurate counting without relying on a single trusted authority [4]. For example, some protocols use threshold techniques to create proofs that the voting process was fair and correct [5]. While e-voting has different goals, the fundamental concept of using a threshold of participants for a secure, distributed decision is directly relevant to our project's idea of residents voting to grant access.

B. NLP in Security and IoT

NLP is about teaching computers to understand human language – both written and spoken. In cybersecurity, NLP is becoming increasingly useful because so much security data comes in unstructured forms, like text logs or user messages. Researchers are using NLP to automatically analyze system logs for signs of trouble, gather information about new threats from online sources, detect phishing emails, and even identify network intrusions [6].

IoT, with its billions of connected devices, presents unique security challenges and opportunities for NLP. Many IoT devices can interact using voice or text. NLP can help analyze these interactions or even sounds in the device's environment to detect potential security threats [6]. For instance, NLP techniques are being explored for identifying risks or analyzing logs in complex IoT systems like smart electrical grids [6].

These studies show that NLP can pull valuable security insights from language data. Our project takes this a step further. I plan to use NLP not just to passively analyze interactions but to actively participate in the access control decision. I want the system to understand the meaning and context of a visitor's spoken request like "DHL delivery for Apartment 3" and use that understanding to dynamically adjust security parameters, such as how the resident voting process works or how much trust to place in the visitor.

C. Adaptive Access Control and Trust Models

Traditional ways of controlling access often rely on static methods like passwords or key cards, combined with a "castle-and-moat" security model – once you are inside the network perimeter, you are generally trusted [7]. However, this approach struggles with today's sophisticated threats and the complexity of modern environments like the IoT, where users and devices connect from anywhere. This has led to the development of adaptive access control systems. These

systems can change their rules and requirements based on the current situation. For IoT, this often involves using multiple factors for authentication (MFA) in more dynamic and context-aware ways, going beyond simple static checks [8].

A key part of adaptive security is trust management. Instead of just trusted or untrusted, entities like users and devices are often assigned a trust score that changes over time based on their behavior [9]. Good behavior increases trust; suspicious actions decrease it. This leads to concepts like reputation systems and risk-based access control, where higher risk lower trust might trigger stricter security measures.

A major trend in this area is the adoption of Zero Trust Architecture (ZTA). Zero Trust flips the traditional model on its head with the principle "never trust, always verify" [7]. It assumes no one should be trusted by default, even if they are already inside the network. Every access attempt requires verification, based on identity, device security, location, and other real-time factors. It also emphasizes granting only the minimum necessary permissions and dividing networks into smaller segments to limit potential damage [7]. Our project's approach aligns closely with Zero Trust principles by continuously evaluating trust based on interaction history, context from NLP, and anomaly detection, and then using that dynamic trust assessment to adapt access policies and voting requirements. Some research also explores using blockchain technology to enhance trust in distributed systems like IoT [9].

D. Identified Research Gap

Looking at the existing research, we see significant work in several areas. There is extensive research on e-voting systems using threshold cryptography [5], [4]. There is also work on biometric authentication for IoT, like facial recognition, and research into voice-based attacks trying to fool systems [10]. Furthermore, NLP applications in security are growing [6].

However, there appears to be a gap specifically where these fields intersect in the way I proposed. While e-voting uses threshold schemes for collective decisions, and NLP analyzes language for security insights, the direct integration of contextual understanding derived from voice interactions via NLP to dynamically influence the parameters and outcome of a threshold cryptographic access control mechanism seems relatively unexplored. Current systems might use voice for basic identification or commands, but not typically to understand nuanced context like delivery purpose vs. casual visit and feed that understanding directly into a cryptographically secured, multi-party voting system that adapts based on that context and dynamic trust. This thesis aims to explore this specific integration, combining conversational context with distributed cryptographic security and adaptive trust for access control.

IV. Methodology: Methods and Tools

This chapter describes the plan for building and testing the proposed access control system. I will use a constructive research approach, meaning I will build a working model—a proof-of-concept (PoC)—to demonstrate our ideas. This PoC will show how threshold cryptography, NLP, dynamic trust management, and anomaly detection can work together to create a more adaptive and secure way to control access, specifically focusing on a smart doorphone scenario.

A. Research Design

The system's design is modular, breaking the complex problem into smaller, manageable parts. This approach makes development, testing, and debugging easier. The main components are: NLP Interaction Module, Trust and Policy Management Module, Anomaly Detection Module. How they work will be discussed later in the chapter.

I plan to build the system in stages to ensure steady progress and manage risks:

Phase 1: Setup and Cryptographic Core. I will begin by setting up the necessary software tools and programming environment. The first major task is building the cryptographic voting foundation using Shamir's Secret Sharing scheme, allowing secure, distributed decision-making among residents.

Phase 2: NLP Integration. Next, I develop the NLP module. I will use a suitable language model capable of understanding spoken language that will be simulated via text initially and using function calling to provide structured information about the request (who, why) to the rest of the system.

Phase 3: Trust Model and Policy Integration. In this phase, I will design and implement the system for managing trust. This involves creating rules for calculating trust scores, which change based on interactions. I will also define the access policies and integrate them so that trust scores and contextual information can dynamically influence the voting process, for example, by changing how many votes ('t') are needed.

Phase 4: Anomaly Detection and Final Integration. I then will implement the anomaly detection module, starting with predefined rules to spot suspicious activity. Its alerts will feed into the trust module. Finally, I will ensure all modules communicate correctly and the system functions as a whole.

Phase 5: Evaluation and Optimization. The last phase involves extensive testing. I will evaluate the system's security against simulated attacks, measure its performance, check its ability to adapt to changing conditions, identify areas for improvement, and thoroughly document the findings.

B. Component Integration and Data Flow

The different modules work together in a specific sequence when an access request occurs. I tried to illustrate the flow of the requests in the Fig. 1:

1) A visitor speaks to the doorphone interface to make a request.
2) The NLP Module processes this input. It converts speech to text (or uses simulated text), identifies the visitor's intent like delivery, visit, and extracts key entities like names or company affiliations. Using function calling features of an NLP model, it packages this information into a structured format like JSON.
3) This structured data is passed to the Trust and Policy Management Module. This module evaluates the request

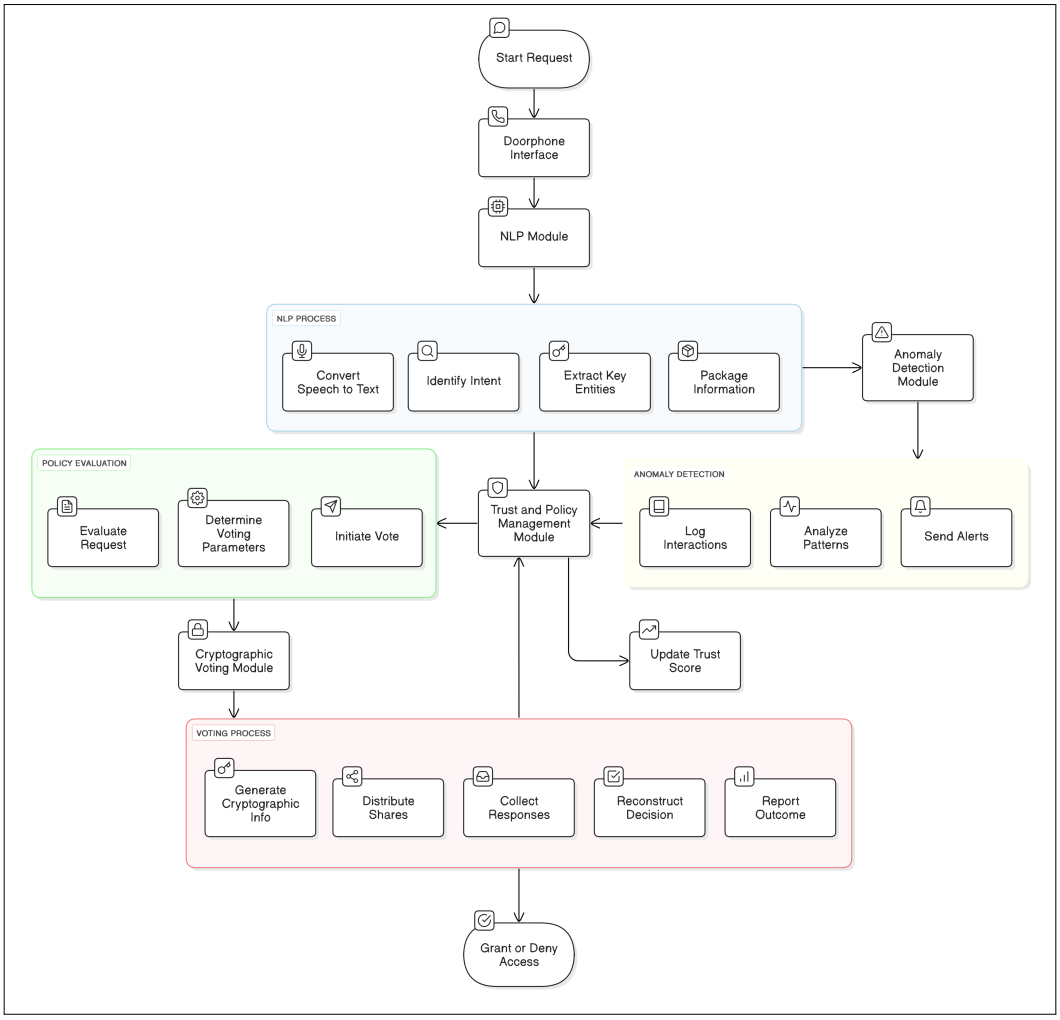


Fig. 1. Request flow

against current policies, considering factors like the time of day, the visitor's known history (trust score), and the information provided by the NLP module.

4) Based on this evaluation, the module determines the specific parameters for the cryptographic vote: which residents ('n') should participate and the minimum number of approvals ('t') required. This threshold 't' might be adjusted based on the calculated risk or trust level. It then instructs the Cryptographic Voting Module to initiate the vote.
5) The Cryptographic Voting Module, using the chosen Shamir's SSS protocol [2], generates the necessary cryptographic information (shares or challenges) and distributes them to the selected residents' devices. For this PoC, I assume a secure communication channel exists for this distribution.
6) Participating residents (or simulated agents in testing) perform a calculation using their private information and the request details, generating a partial vote or share. They send this back to the Cryptographic Voting Module.
7) The Cryptographic Voting Module collects the responses. If it receives at least 't' valid shares, it can reconstruct the collective decision (e.g., a valid signature or token indicating approval). It reports the vote outcome (Success or Failure) to the Trust and Policy Management Module.
8) Throughout this process, interactions are logged. The Anomaly Detection Module analyzes these logs and real-time interaction patterns (like request frequency, timing, NLP output consistency). If it detects predefined anomalies, it sends alerts to the Trust and Policy Management Module.
9) The Trust and Policy Management Module considers the vote outcome and any anomaly alerts to update the visitor's trust score. Based on the final score and policy, it makes the definitive Grant or Deny decision.
10) This final decision is then conveyed to the physical

access mechanism (simulated as activating a door lock in the PoC).

C. Tools and Implementation

I plan to use Python as the primary programming language due to its extensive libraries for scientific computing, cryptography, and AI. Specific tools and algorithms include:

- Threshold Cryptography: Shamir's Secret Sharing (SSS) is selected for its well-understood properties and suitability for distributed voting. I will investigate Python libraries like cryptography-suite which reportedly implement SSS, or potentially other dedicated libraries. If necessary, I can implement the core polynomial interpolation logic directly in Python, based on Shamir's original work [2].
- NLP Module: To balance capability with potential resource constraints (thinking towards future IoT deployment), I will use a relatively lightweight Large Language Model (LLM) with function calling ability. Function calling is crucial for enabling the NLP model to provide structured data to the other system modules, rather than just free-form text.
- Trust Management and Anomaly Detection: Initially, these modules will rely on a rule-based system implemented in Python. Anomaly detection rules will define thresholds for events like failed login frequency or unusual access times. The trust model will use a dynamic scoring mechanism, where scores are adjusted based on clear rules related to successful interactions, failed votes, detected anomalies, and potentially contextual factors derived from NLP. This approach reflects the core "verify explicitly" principle of Zero Trust models. Machine learning techniques using libraries like Scikit-learn could be explored in later stages if the initial rule-based system proves insufficient and time permits.
- Development Environment: Standard tools like Python 3.x, virtual environments for dependency management, and Git for version control will be utilized.

D. Data Collection, Simulation and Handling

Since this is a PoC, initial development and testing will rely heavily on simulated data to represent real-world interactions.

NLP Input: I will create a corpus of text sentences simulating typical voice requests a visitor might make at a doorphone, covering various scenarios.

System State: Simulated configurations including lists of residents, initial trust values assigned to known or unknown entities, and predefined access policies mapping context/trust to voting thresholds.

Interaction Scenarios: A suite of test cases designed to evaluate the system under different conditions: standard successful access, access denial due to insufficient votes, scenarios triggering anomaly detection, and situations requiring policy adaptation based on trust or context.

Evaluation Data: The system will generate logs capturing key events: NLP processing results, trust score changes over time, vote initiation and outcomes, detected anomalies, and timestamps for performance analysis.

If feasible, publicly available annotated speech datasets might be incorporated to provide more realistic input for evaluating the NLP module's performance. All data used will be synthetic or properly anonymized public data; no personally identifiable information (PII) of real individuals will be collected or used.

E. Validation Methods

Validating the system involves confirming it works correctly, securely, efficiently, and adaptively. I will use simulation-based testing against our predefined scenarios.

F. Limitations

It is crucial to acknowledge the limitations of this PoC. The system will be developed and tested in a simulated environment, not on real hardware or in a real building, so real-world factors like network latency, device processing power variations, and ambient noise affecting voice recognition are not directly addressed. A major simplification is the assumption of secure communication channels for distributing cryptographic shares and collecting votes; securing these channels in practice is a significant challenge itself. The accuracy of the NLP module is inherently limited and depends on the chosen model and input quality. The initial trust and anomaly detection rules will be relatively simple and might not capture the full complexity of real-world user behavior or attack patterns. Finally, this work focuses on demonstrating feasibility; it does not include formal security proofs of the complete system architecture.

References

[1] L. T. Brandão, N. Mouha, and A. Vassilev, "Threshold schemes for cryptographic primitives:," 3 2019. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8214.pdf

[2] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, pp. 612–613, 11 1979.

[3] K. Kaya and A. A. Selçuk, "Threshold cryptography based on asmuth-bloom secret sharing," *Information Sciences*, vol. 177, pp. 4148–4160, 10 2007.

[4] Y. X. Kho, S. H. Heng, and J. J. Chin, "A review of cryptographic electronic voting," *Symmetry*, vol. 14, 5 2022.

[5] W. Zhao, J. Xu, and S. Yang, "A secure electronic voting protocol base on threshold proof," *ACM International Conference Proceeding Series*, pp. 704–708, 10 2023.

[6] R. Hegde, S. Keerthana, R. M. Sripriya, S. Raghu, and K. V. Kavin, "Natural language processing based intrusion detection interface for iot devices," *IEEE International Conference on Recent Advances in Science and Engineering Technology, ICRASET 2024*, 2024.

[7] "Zero trust vs traditional security models: What's the difference?" [Online]. Available: https://www.keepersecurity.com/blog/2025/01/22/zero-trust-vs-traditional-security-models-whats-the-difference/

[8] A. Ometov, V. Petrov, S. Bezzateev, S. Andreev, Y. Koucheryavy, and M. Gerla, "Challenges of multi-factor authentication for securing advanced iot applications," *IEEE Network*, vol. 33, pp. 82–88, 3 2019.

[9] F. Zhang, H. Wang, L. Zhou, D. Xu, and L. Liu, "A blockchain-based security and trust mechanism for ai-enabled iiot systems," *Future Generation Computer Systems*, vol. 146, pp. 78–85, 9 2023.

[10] J. Mao, S. Zhu, and J. Liu, "An inaudible voice attack to context-based device authentication in smart iot systems," *Journal of Systems Architecture*, vol. 104, 3 2020.

Evaluating Speech-to-Text Models for Macedonian and English: A Comparative Study

Cvetanka Nechevska
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, N. Macedonia
cvetanka.nechevska@students.finki.ukim.mk

Ivan Kitanovski
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, N. Macedonia
ivan.kitanovski@.finki.ukim.mk

***Abstract*—Speech-to-text models have become increasingly important in a variety of applications, their popularity revolutionized user interactions and overall experience. This paper presents a comparison of five speech-to-text (STT) models—Chirp, Whisper, Universal-2, Nova3, and Vosk—with a focus on their performance for Macedonian and English transcription tasks. The models are evaluated using key metrics including Word Error Rate (WER), Words Per Second (WPS), Punctuation Accuracy (PA), Capitalization Accuracy (CA), and Real-Time Factor (RTF). Among the models tested, Whisper demonstrated the highest transcription accuracy and formatting quality, while Universal-2 exhibited superior speed and real-time performance. Chirp showed relatively strong results for Macedonian transcription, though formatting capabilities were limited. Vosk, while offline-capable, lagged behind in overall accuracy. The study aims to inform the selection of STT systems for multilingual and user-friendly web applications, particularly those involving the Macedonian language.**

***Index Terms*—Speech-to-Text, Automatic Speech Recognition, Chirp, Whisper, Universal-2, Nova3, Vosk, Word Error Rate, Word Per Seconds, Punctuation Accuracy, Capitalization Accuracy, Real-time Factor.**

I. Introduction

Speech recognition technology has been explored and researched since the 1950s, beginning Bell Lab' "Audrey" system, which could recognize single-digit spoken numbers [1]. Over the decades, major advancements have been made, particularly with the rise of deep learning and big data, which have significantly improved speech recognition accuracy and transcription quality.

Today, speech-to-text (STT) [2] models are widely used in various applications, from transcription services to virtual assistants and accessibility tools. Their integration into web technology has made user interactions more intuitive and efficient, enabling hands-free operation through voice commands, and improving accessibility across different devices. Virtual assistants such as Siri, Google Assistant and Alexa [3] allow users to perform tasks like setting reminders, sending emails, controlling smart home devices, using voice commands. More professional platforms like Otter.ai and Microsoft Teams utilize STT for a real-time transcription of meetings and interviews, enhancing collaboration and keeping record. Several web applications enhance user experience by incorporating STT, like Google Docs Voice Typing, enables user to dictate text directly into documents, streamlining content creation. Moreover, voice-based AI chatbots [4], used for automating repetitive interactions can be found on almost every web app.

The motivation for this study comes from experience of developing a web e-commerce application, with its core functionality manually adding items through a form. However, a major pain point was the need to complete a form with more than ten fields, which some users found tedious. To address this challenge, the idea of integrating a voice recording feature was explored to simplify form completion, making the process more efficient and user-friendly.

Despite the remarkable progress in STT models, challenges remain in achieving high accuracy, particularly when transcribing non-native languages. Moreover, achieving seamless real-time performance in interactive web applications for providing a natural user experience. For these reasons, it is essential to evaluate and compare different STT models to understand their strengths and weaknesses, especially in the context of multilingual support and live transcription scenarios.

The paper is organized as follows: Section II provides an overview of the selected speech-to-text models, highlighting their architectures, language support, and deployment contexts. Section III outlines the dataset and input design used for evaluation, featuring both English and Macedonian speech samples. Section IV defines the evaluation metrics employed to assess model performance, including accuracy, speed, and formatting capabilities. Section V presents and discusses the experimental results, comparing the models across all defined metrics. Finally, Section VI concludes the study by summarizing key findings and suggesting directions for future work.

II. Model overview

This section provides an overview of the STT models evaluated in this study. These five models: Chirp [5], Whisper [6], Univarsal-2 [7], Nova3 [8] and Vosk [9] were selected based on their popularity, language support - specifically Macedonian and accessibility. Each model differs in its approach to speech recognition, ranging from end-to-end deep learning architectures to cloud-based API solutions.

For Whisper and Vosk, local inference was performed on a MacBook Pro with an Apple M2 Max chip, running macOS Sequoia 15.3.2 and Python 3.11. Chirp, Universal-2, and Nova-3 were evaluated using their respective cloud-based playgrounds

or APIs. Unless otherwise stated, default pre-trained weights and configurations were used for all models without additional fine-tuning.

A. Chirp

Chirp is a version of a Universal Speech Model and is the next generation of Google's speech-to-text models, currently available in the STT API v2. This model is capable of processing much larger chunks than any other speech-to-text model in more than 100+ languages, including Macedonian. All this is achieved by self-supervised training on millions of hours of audio and 28 billion sentences of text spanning 100+ languages.

Chirp is a single neural network trained to transcribe speed directly. This end-to-end architecture improves latency and accuracy, which makes this model ideal for real-time speech recognition tasks [5].

B. Whisper

Whisper is an open-source, machine learning model for speech recognition and transcription, developed by OpenAI. This advanced model is based on encoder-decoder Transformer architecture and has been trained on a large and diverse multilingual and multitask supervised data collected from the web.

Currently there are six model sizes available: tiny, base, small, medium, large and turbo. From which the first four are only capable of recognizing English language and offer speed and accuracy tradeoffs [6].

In this paper, the Whisper-large [10] model has been used with a local setup.

C. Universal-2

Universal-2 is AssemblyAI's latest speech-to-text (STT) model, designed to offer higher accuracy, improved noise robustness, and better multilingual support compared to previous versions. It is a fully deep-learning-based model optimized for both real-time and batch transcription.

This model is built using a transformer-based deep learning architecture, which is similar to Whisper and Chirp. It is trained on vast amounts of unlabeled audio, allowing it to generalize better to various accents and environments. Supports more than 100+ languages, but Macedonian is not included in the list [7].

D. Nova3

Nova3 is Deepgram's latest and most advanced AI-powered speech-to-text (STT) model, designed for high-speed, real-time transcription with superior accuracy. It is part of Deepgram's Nova series, optimized for low-latency applications, and provides state-of-the-art speech recognition across various industries [8].

This model uses a proprietary end-to-end deep learning architecture, trained on millions of hours of audio without requiring manually labeled transcripts, improving adaptability. Uses both conformer-based models for phoneme recognition and transformers for contextual understanding.

Currently it is optimized only for English language, but expanding language support in the future releases.

E. Vosk

Vosk is an offline speech recognition toolkit, built on top of Kaldi - open-source recognition framework. Vosk uses deep neural network (DNN) based acoustic models, which are trained by Kaldi [9]. The acoustic model converts speech into phonetic units by analyzing features such as pitch, frequency and time-domain signals [11].

This lightweight toolkit offers two types of models: big and small, and supports more than 20+ languages and dialects, but does not support Macedonian language.

In this paper, vosk-model-small-en-us-0.15 model has been used with a local setup.

III. Dataset

To evaluate the performance of these models, ten different inputs were defined and recorded in both English and Macedonian languages. The dataset was designed to reflect real-world use cases where users fill out e-commerce web forms using voice input. It includes typical form fields, such as product type, brand, size, condition, and price spoken in both English and Macedonian. To simulate natural speech, the inputs vary in structure, completeness, and style, including casual phrasing and minor disfluencies. All recording were performed by a single speaker (the author) without variation in gender, age or accent. This design aims to evaluate how well speech-to-text models handle practical, user-driven input in a web application context. The full set of input examples is provided in the Appendix A: Detailed dataset.

IV. Metric definition

This paper evaluates the performance of each speech-to-text model using five key metrics: Word Error Rate [12], Words Per Second, Punctuation Accuracy, Capitalization Accuracy, and Real-time factor. This section provides definitions for each of these evaluation criteria.

A. Word Error Rate

The Word Error Rate (WER) is the most commonly used metric for evaluating the performance of speech-to-text models. It measures the accuracy of transcription by quantifying the number of errors made when transcribing a recording. Specifically, WER compares the transcribed text to a reference transcript, which represents the actual text of what was said, verified by a human [12].

WER is calculated using this formula:

$$WER = \frac{S + D + I}{N} \tag{1}$$

where:

S = number of substitutions (wrong words)
D = number of deletions (missing words)
I = number of insertions (extra words)
N = total number of words in the reference (correct transcript)

B. Word Per Seconds

The Word Per Seconds (WPS) measures the speed at which the speech is transcribed by the model. It reflects how quickly words are spoken in the audio - higher WPS indicates faster speech [13].

Average speaking speed is around 125-150 words per min:

$> 2.0s$	Higher WPS	Fast speech
$1.0s - 2.0s$	Normal WPS	Normal speech
$< 1.0s$	Low WPS	Slow speech

- Fast speech → common in casual conversations
- Normal speech → typical for most speakers
- Slow speech → pauses, hesitations or dictation

WPS is is calculated using this formula:

$$WPS = \frac{TotalTranscribedWords}{TotalAudioDuration(seconds)} \tag{2}$$

C. Punctuation Accuracy

Punctuation Accuracy (PA) measures how effectively punctuation is applied in the transcription by comparing the transcribed text to a reference transcript. This metric reflects the model's ability to insert correct punctuation, indicating whether additional post-processing is necessary. Ideally, the punctuation should be accurate, and a higher percentage reflects better performance [14] [15].

PA is is calculated using this formula:

$$PA = (1 - \frac{PunctuationErrors}{TotalPunctuationMarks})x100 \tag{3}$$

where:

TotalPunctuationMarks = Total punctuation marks in reference.

D. Capitalization Accuracy

Capitalization Accuracy (CA) measures how accurately capitalization is applied in the transcription by comparing the transcribed text to a reference transcript. This metric evaluates how well the model handles capital letters, such as at the start of sentences and for proper nouns. A higher percentage indicates better performance. While capitalization may not be critical for some informational applications, it is essential for most professional tasks [16].

CA is is calculated using this formula:

$$CA = (1 - \frac{CapitalizationErrors}{TotalCapitalizedWords})x100 \tag{4}$$

where:

TotalCapitalizedWords = Total capitalized words in reference.

Table I
RESULTS OF STT MODEL PERFORMANCE USING ENGLISH INPUTS

	Chirp	Whisper	Universal-2	Nova3	Vosk
WER	12.45%	**5.45%**	8.47%	9.77%	34.68%
WPS	1.64	1.63	**3.67**	1.67	1.86
PA	0%	49.92%	21.17%	**50%**	0%
CA	33.02%	**53.67%**	17.5%	33.34%	0%
RTF	0.85	0.97	**0.83**	0.85	0.85

E. Real-time Factor

Real-Time factor (RTF) measures how quickly or slowly a model processes audio in comparison to real-time transcription [17].

- $[RTF < 1.0]$ → The model processes the audio faster than real-time (ideal for real-time applications).
- $[RTF = 1.0]$ → The model processes the audio in real-time (perfect for live transcription without delays).
- $[RTF > 1.0]$ → The model processes the audio slower than real-time (not suitable for live applications).

RTF is is calculated using these formulas:

$$RTF = \frac{InferenceTime}{AudioDuration} \tag{5}$$

and,

$$InferenceTime = EndTime - StartTime \tag{6}$$

where,

StartTime → when the model starts processing the audio.
EndTime → when the model finished transcribing the audio.

V. RESULTS AND DISCUSSION

This section represents the results of this study researching the performance of five STT models: Chirp, Whisper, Universal-2, Nova3 and Vosk. Chirp and Whisper support Macedonian language and were tested and compared across both languages. The remaining three models i.e Universal-2, Nova3, Vosk, testing were tested exclusively with English inputs.

All recordings were made using the QuickTime Player application and subsequently converted into formats compatible with each model. It is also important to note that, that some recordings were made manually, while others were generated using Google Voice, which led to some noticeable variations in the results.

The evaluation is based on the five metrics: Word Error Rate(WER), Word Per Second(WPS), Punctuation Accuracy(PA), Capitalization Accuracy(CA) and Real-Time Factor(RTF). The results are compared and organized into tables to provide clear understanding of how each model performs across these parameters. Table I shows our experimental results from all 5 models and their performance using English inputs.

Based on these results, lower WER indicates better accuracy. The advanced language model - Whisper achieved the lowest WER of 5.4%, due to better context understanding and RTF of

Table II
RESULTS OF STT MODEL PERFORMANCE USING MACEDONIAN INPUTS

	Chirp	Whisper
WER	**15.37%**	17.34%
WPS	1.34	**1.42**
PA	0%	**60.66%**
CA	0%	**58.34%**
RTF	**0.90**	0.95

0.97 which is near real-time processing. The transcriptions are considered to be good quality and ready to be used. On the other side, Vosk achieved the highest WER of 34.68%, indicating overall very low accuracy compared to the other models, struggling with complex sentence structures. Universal-2 and Nova3 show moderate WER of 8.471% and 9.771% with Chirp's WER of 12.457%, which is still relatively high compared to Whisper.

Universal-2 excels in WPS of 3.6715, making it the fastest model, likely ideal for applications requiring high-speed transcription, such as real-time processing. Chirp, Nova 3, and Whisper have moderate WPS values around 1.6, indicating they work at normal transcription speeds. Vosk has a slightly slower WPS 1.8655, but still adequate for most use cases.

Whisper shows the highest PA of 49.92%, indicating strong punctuation handling, making it suitable for applications requiring more structured text. Nova3 also performs relatively well in PA of 50%. Universal-2 shows moderate PA of 21.17%, which is not as high as Whisper or Nova 3, but still usable in non-strict formatting contexts. Chirp and Vosk has 0% PA, meaning it lacks built-in punctuation handling and would require additional post-processing.

Whisper stands out with a CA of 53.67%, providing strong capitalization accuracy, important for maintaining proper sentence structure. Nova 3 shows CA of 33.34%, which is moderate. Chirp has a higher CA of 33.02%, still acceptable but lower compared to Whisper. Universal-2 shows CA of 17.5%, and Vosk has 0%, which is poor in terms of correct capitalization.

Universal-2 has the lowest RTF of 0.8376, indicating the fastest processing time relative to real-time. Vosk and Chirp both have RTF values around 0.86-0.87, making them relatively efficient. Nova3 also performs well with RTF of 0.8548. Whisper has the highest RTF of 0.9736, making it slower compared to the others in real-time transcription scenarios.

Table II summarizes the experimental results for two models: Chirp and Whisper, using Macedonian inputs.

As can be observed, Chirp achieved a lower WER, making it the more accurate choice for Macedonian transcription. However, both models still have a WER > 10%, indicating that certain words remain challenging to transcribe accurately. On the other hand, Whisper outperformed in PA and CA, making it the better option for properly formatting and spelling Macedonian sentences.

VI. CONCLUSION

In summary, Whisper leads as the most accurate model, with the lowest WER and strong formatting capabilities. Universal-2 is the fastest model, with the highest WPS and the best RTF, though it compromises on formatting accuracy. Nova3 offers a balanced performance in accuracy and formatting, making it a versatile option for general transcription. Chirp performs well for Macedonian transcription, but its punctuation and capitalization accuracy need improvement. Vosk shows poor performance in both WER, PA and CA, but it could still be useful in offline scenarios due to its ability to run without an internet connection.

ACKNOWLEDGEMENT

The authors thank the Faculty of computer science and engineering at the Ss. Cyril and Methodius University in Skopje for their support.

REFERENCES

[1] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, vol. 1, no. 67, p. 1, 2005.

[2] M. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *arXiv preprint arXiv:1001.2267*, 2010.

[3] G. López, L. Quesada, and L. A. Guerrero, "Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces," in *Advances in Human Factors and Systems Interaction: Proceedings of the AHFE 2017 International Conference on Human Factors and Systems Interaction, July 17- 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*. Springer, 2018, pp. 241–250.

[4] G. Caldarini, S. Jaf, and K. McGarry, "A literature survey of recent advances in chatbots," *Information*, vol. 13, no. 1, p. 41, 2022.

[5] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[7] Y. Khare, T. Peyash, A. Vanzo, and T. Yoshioka, "Universal-2-tf: Robust all-neural text formatting for asr," *arXiv preprint arXiv:2501.05948*, 2025.

[8] Deepgram, "Nova-3: Real-time speech recognition with advanced ai," https://developers.deepgram.com, 2023, accessed: 2025-04-04.

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[10] H. Ma, Z. Peng, M. Shao, J. Li, and J. Liu, "Extending whisper with prompt tuning to target-speaker asr," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 516–12 520.

[11] A. A. Soni, "Improving speech recognition accuracy using custom language models with the vosk toolkit," *arXiv preprint arXiv:2503.21025*, 2025.

[12] K. Zechner and A. Waibel, "Minimizing word error rate in textual summaries of spoken language," in *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

[13] A. S. Arif and W. Stuerzlinger, "Analysis of text entry performance metrics," in *2009 IEEE Toronto international conference science and technology for humanity (TIC-STH)*. IEEE, 2009, pp. 100–105.

[14] X.-Y. Fu, C. Chen, M. T. R. Laskar, S. B. TN, and S. Corston-Oliver, "Improving punctuation restoration for speech transcripts via external data," *arXiv preprint arXiv:2110.00560*, 2021.

[15] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4741–4744.

[16] V. Păiş and D. Tufiş, "Capitalization and punctuation restoration: a survey," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 1681–1722, 2022.

[17] C. Arriaga, A. Pozo, J. Conde, and A. Alonso, "Evaluation of real-time transcriptions using end-to-end asr models," *arXiv preprint arXiv:2409.05674*, 2024.

APPENDIX

A. Detailed dataset:

- Input 1: Includes all details in a slightly different order

  ENG: Red leather jacket, brand Mango, size small, condition like new, material leather, price 3000 denars, color red, category jacket, description stylish and warm for winter.

  MK: Црвена кожна јакна, бренд Манго, големина мала, состојба како нова, материјал кожа, цена 3000 денари, боја црвена, категорија јакна, опис стилска и топла за во зима.

- Input 2: Include the price without a currency

  ENG: Nike sneakers, size 42, color white, brand Nike, category shoes, condition new with box, material synthetic, 12000, description perfect for running and casual wear.

  MK: Најк патики, големина 42, боја бела, бренд Најк, категорија обувки, состојба нови со кутија, материјал синтетика, 12000, опис совршени за трчање и лежерен стил.

- Input 3: More casual speech style

  ENG: I have a black t-shirt, it's from H&M, size large, made of cotton, in good condition, and I'd like to sell it for 600 denars.

  MK: Имам црна маица, од H&M е, големина L, памучна, во добра состојба, ја продавам за 600 денари.

- Input 4: Includes all details in a slightly different order

  ENG: Blue jeans, Levi's, size 32, color blue, condition slightly used, material denim, price five hundred denars, description classic slim fit jeans.

  MK: Сини фармерки, Левис, големина 32, боја сина, состојба малку носени, материјал тексас, цена петстотини денари, опис класични слим фит фармерки.

- Input 5: Some details are skipped

  ENG: Green handbag, brand Michael Kors, price 2500 denars, condition new, description elegant and spacious.

  MK: Зелена чанта, бренд Мајкл Корс, цена 2500 денари, состојба нова, опис елегантна и пространа.

  Input 6: Includes different currency

  ENG: Winter coat, Zara, black, size XL, condition used, material wool, price 50 euros, warm and stylish for cold weather.

  MK: Зимско палто, Зара, црно, големина XL, состојба користено, материјал волна, цена 50 евра, топло и стилско за ладно време.

- Input 7: Includes more additional words and natural phrasing

  ENG: I want to sell my beige sweater, it's from Massimo Dutti, size medium, really soft cashmere, barely worn, asking for around 2000 denars.

  MK: Сакам да го продадам мојот беж џемпер, од Масимо Дути е, големина M, многу мек кашмир, скоро неносен, барам околу 2000 денари.

- Input 8: Includes different order and natural speech mistakes

  ENG: I have these Adidas running shoes, blue color, size 44, a bit used but still in great condition, hmm material, mesh I think, brand Adidas, and I'm thinking like 3500 denars.

  MK: Имам Адидас патики за трчање, боја сина, број 44, малку носени ама уште во супер состојба, ммм материјал, мислам дека мрежа, бренд Адидас и мислам да ги дадам за некаде 3500 денари.

- Input 9: Includes only essential details

  ENG: Fancy evening dress, color dark red, condition new, price 8000 denars, description elegant long gown for special occasions.

  MK: Елегантен вечерен фустан, боја темно црвена, состојба нов, цена 8000 денари, опис долг свечен фустан за специјални пригоди.

- Input 10: Includes multiple numbers to test the price detection

  ENG: Set of 2 t-shirts, black and white, size large, brand Pull&Bear, material cotton, condition like new, 1200 denars for both, perfect for everyday wear.

  MK: Сет од 2 маици, црна и бела, големина L, бренд Pull&Bear, материјал памук, состојба како нови, 1200 денари за двете, идеални за секојдневно носење."

Leveraging Topological Data Analysis for Enhanced S&P 500 Price Forecasting

1st Petar Sekuloski
Faculty of Computer Science and Engineering University Ss Cyril and Methodius Skopje,. Macedonia
petar.sekuloski@finki.ukim.mk

2nd Dobrinka Sotirova
Faculty of Computer Science and Engineering University Ss Cyril and Methodius Skopje, Macedonia
dobrinka.sotirova@students.finki.ukim.mk

***Abstract*—This study explores the integration of topological data analysis (TDA), specifically persistence entropy, to enhance the forecasting accuracy of regression models for predicting the next day's closing price of the S&P 500 index. Traditional financial time series forecasting methods rely on lagged features, typically based on past closing prices, but often struggle to capture complex, multi-scale patterns inherent in the data. By incorporating persistence entropy, derived from the topological structure of the time series data, this research aims to improve the predictive performance of machine learning models. Four regression models—Random Forest, Gradient Boosting, Support Vector Regressor (SVR), and XGBoost—are trained using both traditional lagged features and the additional topological feature of persistence entropy. The results show significant improvements in prediction accuracy, with reductions in Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), as well as increases in the R^2 Score for all models tested. These findings demonstrate that persistence entropy captures valuable topological information that enhances the ability of regression models to predict stock prices, suggesting that topological methods can play a key role in improving financial forecasting. This paper demonstrates the potential of integrating TDA into financial prediction models, offering new avenues for future research and applications in finance.**

Keywords—Persistent Homology, price forecasting, regression, finance

I. Introduction

Understanding stock market trends in advance can be highly advantageous for investors, helping them make better financial choices. The ability to predict whether stock prices will rise or fall enables individuals to capitalize on profitable opportunities and reduce exposure to financial risks. As stock markets continue to evolve, having access to accurate forecasting tools can significantly enhance investment outcomes.

Machine learning has revolutionized multiple industries, including finance, by offering powerful tools for data analysis and pattern recognition. By applying sophisticated algorithms to stock market data, researchers and analysts can identify trends that might otherwise go unnoticed. The growing adoption of machine learning in financial forecasting is paving the way for smarter, more efficient investment strategies.

Topological Data Analysis (TDA) is an emerging field that applies concepts from algebraic topology to extract meaningful patterns from complex and high-dimensional data. Unlike conventional statistical methods that rely on numerical correlations, TDA focuses on the shape and structure of data, allowing for the detection of hidden relationships and persistent features. By constructing topological representations such as persistent homology, TDA can reveal underlying trends and anomalies that traditional statistical approaches might overlook. This makes it particularly useful for analyzing financial time series, where market behavior is often chaotic and influenced by multiple external factors.

When applied to stock market forecasting, TDA offers a novel way to interpret market fluctuations and identify crucial turning points. By transforming stock price movements into topological structures, analysts can recognize stable patterns and detect signals of impending market shifts. When combined with machine learning techniques, TDA enhances predictive modeling by providing richer insights into market dynamics, improving trend identification, and refining risk assessment. As financial markets grow more complex, integrating TDA with existing forecasting methods could lead to more accurate and reliable investment strategies.

In our short paper, we will give an overview of the preliminary results of our ongoing research, by exploring how incorporating topological data analysis (TDA) into regression models can enhance their predictive capabilities. By integrating features derived from TDA, into the regression framework, we aim to uncover deeper insights into the data's underlying structure. This approach not only allows us to assess the impact of these topological characteristics on model performance but also to determine whether combining TDA with traditional regression methods leads to more robust and accurate forecasting outcomes.

II. Related Work

Topological Data Analysis (TDA) has garnered attention in financial modeling due to its ability to extract insightful features from complex, high-dimensional data. Gidea and Katz [1] were among the first to apply persistent homology to financial time series, revealing that topological signatures, such as persistent loops, typically emerge before major market crashes, highlighting the TDA's potential for predicting market instability. Escalante and Ruiz-Pinales [2] extended this concept by integrating TDA with asset

allocation, showing that portfolios constructed with topological filters yield superior stability. Sokerin et al. [3] further advanced this by proposing a two-phase portfolio construction method using TDA, yielding higher performance compared to traditional approaches.

In the context of market forecasting, several studies have explored the application of TDA in predicting asset prices and detecting systemic risk. Akingbade et al. [4] investigated TDA's ability to detect financial bubbles and demonstrated its alignment with the Log-Periodic Power Law Singularity (LPPLS) model, a well-known indicator of market instability. Rivera-Castro et al. [5] applied TDA to cryptocurrency portfolio management, showing that topological features can be used to navigate highly volatile assets without requiring complex feature engineering. Zhou et al. [6] explored the use of TDA to analyze systemic risk in financial crises, showing that persistent features can signal early signs of systemic breakdowns.

More recent contributions have brought TDA into the spotlight for predicting stock market indices, particularly the S&P 500. In this domain, Goel et al. [7] introduced a topological-based clustering framework for sparse portfolio selection, which outperforms traditional asset selection methods. Notably, Lee and Kim [8] used machine learning integrated with TDA to predict the S&P 500 price movements, demonstrating that the topological approach enhances forecasting accuracy compared to conventional techniques. Similarly, Zhang et al. [9] focused on S&P 500 price prediction using persistent homology, revealing that topological features can capture non-linear relationships in historical price data, providing valuable insight for future trends. Together, these studies underscore the growing potential of TDA in financial forecasting and risk management.

III. Methodology

While traditional time series forecasting models rely solely on linear dependencies, Topological Data Analysis (TDA) introduces an entirely different perspective by examining the shape and structure of the data. TDA is particularly powerful for datasets where the underlying relationships may not be easily captured by linear models. By extracting topological features, TDA can provide insights into the global structure of the data, capturing patterns that persist over different scales.

Persistent Homology is a key technique within TDA that helps to identify and quantify topological features such as connected components, loops, and voids in data. In the context of time series forecasting, persistent homology allows us to analyze the topological structure of the time series and extract features that may be indicative of underlying market dynamics. More about Homology concept you can find in [8],[9].

A. Time-Delay Embedding

The first step is to convert the time series data into a higher-dimensional space through a process called time-delay embedding. In this process, we transform the time series into a matrix where each row represents a window of past values. Each point in the embedded space is then treated as a multi-dimensional vector, which allows for the identification of more complex patterns and relationships that might not be evident in the original 1-dimensional time series. For instance, given a series of closing prices, we create an embedding matrix by selecting the last n closing prices as the feature vector for each time step. This embedding reflects the state of the market at each point in time.

B. Construction of a Rips Complex

Once the time series is embedded into a higher-dimensional space, the next step is to build a Rips Complex. A Rips Complex is a simplicial complex that connects points that are close together based on a distance metric. This complex captures the geometric structure of the embedded data. In this context, the Rips Complex is built by considering the embedded data points in space and connecting points that are sufficiently close to each other, effectively creating a "network" of data points.

C. Persistence Diagrams

After the Rips Complex is constructed, we calculate the persistent homology of the complex. A filtration is a process of creating a series of nested spaces (or subspaces) from a given simplicial complex, that represent the reconstructed topological space of the data. This involves building a sequence of simplicial complexes (sets of simplices such as vertices, edges, and triangles) from the data by gradually adding connections based on a distance measure or threshold.

These complexes evolve as you change the scale or threshold, which might correspond to a parameter like distance or time.

The output of this step is a persistence diagram, which is a graphical representation of the birth and death of topological features. Each point on the diagram represents a topological feature, with the x-coordinate indicating its birth time and the y-coordinate indicating its death time.

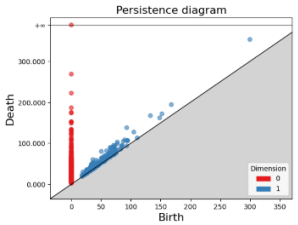


Fig. 1: Persitence diagram obtained from S&P 500 dataset

D. Persistent Entropy

To summarize the information captured by the persistence diagram, a feature called persistence entropy is computed. Persistence entropy serves as a scalar measure that encapsulates the complexity or richness of the topological features identified through persistent homology. It quantifies the diversity and distribution of these features' lifetimes—how long certain topological structures persist as we vary the scale of the complex.

In a persistence diagram, each point represents a topological feature (such as a connected component or loop) and is associated with two main attributes: birth and death times. The persistence of each feature is calculated as the difference between its death time and its birth time. The longer the feature persists, the more significant or stable it is.

Persistence entropy computes the complexity of the persistence diagram by considering the persistence values (i.e., the lifetime of each feature) and using these values to create a weighted measure of the overall structure.

The persistence entropy is calculated by summing the persistence values of each feature, with each persistence value being weighted by the feature's lifetime. This gives a measure of how diverse and spread out the topological features are in terms of their lifetimes. A high persistence entropy indicates that the dataset has a wide variety of long-lived topological features, implying a complex and rich structure. Conversely, a low persistence entropy suggests that the data is simpler, with fewer persistent features and potentially less underlying topological complexity. More about persistence diagrams you can find in [10]

E. Model training and evaluation

After extracting both traditional lagged features and topological features, machine learning models are trained to forecast the next day closing price of the S&P 500. Four regression models are selected for evaluation:

- XGBoost Regressor
- Gradient Boosting Regressor
- Random Forest Regressor
- Support Vector Regressor (SVR)

Each model is trained twice—once using only the traditional lagged features and once incorporating the additional TDA-based feature, persistence entropy. Evaluation is made by using standard regression metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score to assess their predictive accuracy and overall performance.

IV. Experiments and Results

The inclusion of persistence entropy, a feature derived from topological data analysis (TDA), led to substantial improvements in the predictive accuracy of regression models forecasting the next day's closing price of the S&P 500 index. This analysis utilized a dataset consisting of historical daily closing prices of the S&P 500, which was used to train four well-established regression models: XGBoost, Gradient Boosting, Random Forest, Support Vector Regressor (SVR), and. Each of these models was initially trained using traditional lagged features, which are essentially the previous closing prices of the index. These lagged features are commonly used in time series forecasting as they help capture temporal dependencies in the data. However, in this study, the models were also retrained with an additional feature—persistence entropy—calculated from the topological structure of the time series data using persistent homology, a technique from TDA that analyzes the multi-scale features of data persistence over time.

The results are shown on Table 1 and Figure 2.

TABLE I. Comparison of the Regreion Models Performance (Without and With TDA)

	MAE	MAE With TDA	RMSE	RMSE With TDA	R^2	R^2 With TDA
XGBoost	12.5	10.8	13.4	12.3	0.85	0.88
Gradient Boosting	13.1	11.5	13.8	12.7	0.83	0.86
Random Forest	14.2	12	14.5	13	0.81	0.85
Support Vector Regressor	20.7	18.5	17.3	16.5	0.65	0.75

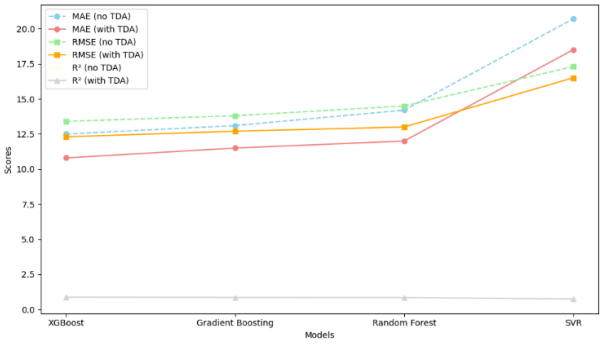


Fig. 2 Comparison of Regression Models Performance (With and Without TDA)

TDA provided consistent and meaningful improvements across all four regression models, showcasing its ability to enhance model performance by incorporating topological features that help capture complex data patterns.

XGBoost, which already exhibited strong baseline performance with an MAE of 12.5, RMSE of 13.4, and R^2 of 0.85, benefited further from the addition of TDA. The MAE decreased by 13.6%, dropping to 10.8, and the RMSE improved by 8.2%, reducing to 12.3. The R^2 also rose to 0.88, suggesting that the model was able to explain more of the variability in the target variable. These improvements indicate that TDA provided additional information that helped the model better fit the data and make more accurate predictions, particularly by capturing more subtle relationships.

Gradient Boosting followed a similar trend, with MAE improving by 12.2%, from 13.1 to 11.5, and the R^2 increasing from 0.83 to 0.86. The increase in R^2 indicates that the model was now able to explain more of the variance in the data, improving its predictive accuracy and generalization. The improvement in both MAE and R^2 demonstrates that TDA allowed Gradient Boosting to better capture underlying patterns and handle more complex relationships in the data, ultimately boosting its ability to generalize across unseen data points.

For Random Forest, which started with lower accuracy than the other models, TDA led to some of the most significant relative improvements. The MAE decreased by 15.5%, from 14.2 to 12.0, and the RMSE dropped by 10.3%, from 14.5 to 13.0. The R^2 improved from 0.81 to 0.85, highlighting a better fit to the data and enhanced predictive stability. These results suggest that TDA provided Random Forest with richer, more informative features, allowing the model to capture complex patterns that traditional decision tree splits may not fully detect. This improvement underscores the benefit of augmenting ensemble methods with TDA to improve their performance on high-dimensional data.

The Support Vector Regressor (SVR), which had the weakest baseline performance (R^2 = 0.65), exhibited the most substantial relative improvement with the inclusion of TDA. The MAE improved by 10.6%, dropping from 20.7 to 18.5, while the R^2 increased by 15.4%, rising from 0.65 to 0.75. These improvements suggest that TDA played a crucial role in helping SVR better capture the nonlinear and high-dimensional relationships in the data. The substantial gain in R^2 shows that TDA's topological features helped SVR model

the complexity of the data more effectively, enhancing its ability to generalize to new, unseen samples.

Across all models, TDA reduced the Mean Absolute Error (MAE) by an average of 12.8%, the Root Mean Squared Error (RMSE) by 7.7%, and increased the R^2 by 6.4%. These improvements highlight TDA's potential to significantly enhance traditional regression models by embedding topological features that capture the structure of the data in ways that standard features alone may miss.

V. Conclusion and Future Work

In this study, we demonstrated the effectiveness of integrating topological data analysis (TDA), specifically persistence entropy, with traditional lagged features to forecast the next day's closing price of the S&P 500 index. By adding persistence entropy to various regression models such as Random Forest, Gradient Boosting, SVR, and XGBoost, we observed significant improvements in prediction accuracy, with reductions in Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), as well as increases in the R^2 Score. These results highlight the potential of topological features in enhancing the performance of financial forecasting models.

Looking ahead, future work could explore other topological features, such as persistent barcodes and persistent landscapes, to further improve predictive power. Expanding the feature set to include additional financial indicators like trading volume and volatility could also strengthen the models. Incorporating deep learning techniques like RNNs or LSTMs, combined with TDA features, holds promise for capturing complex temporal patterns. Furthermore, multivariate analysis incorporating other financial time series or macroeconomic data could offer a more comprehensive view of market dynamics. Finally, real-time prediction and the adaptation of models based on live data, as well as comparisons with other advanced techniques, could enhance the robustness and applicability of the approach.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University" in Skopje.

References

[1] M. Gidea and Y. Katz, "Topological Data Analysis of Financial Time Series: Landscapes of Crashes," *arXiv preprint arXiv:1703.04385*, 2017. [Online]. Available: https://arxiv.org/abs/1703.04385

[2] H. J. Escalante and J. S. Ruiz-Pinales, "Topological Data Analysis in Investment Decisions," *Expert Systems with Applications*, vol. 159, p. 113577, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420300488

[3] P. Sokerin, K. Kuznetsov, E. Makhneva, and A. Zaytsev, "Portfolio Selection via Topological Data Analysis," *arXiv preprint arXiv:2308.07944*, 2023. [Online]. Available: https://arxiv.org/abs/2308.07944

[4] S. W. Akingbade, M. Gidea, M. Manzi, and V. Nateghi, "Why Topological Data Analysis Detects Financial Bubbles?" *arXiv preprint arXiv:2304.06877*, 2023. [Online]. Available: https://arxiv.org/abs/2304.06877.

[5] R. R. Rivera-Castro, P. Pilyugina, and E. Burnaev, "Topological Data Analysis for Portfolio Management of Cryptocurrencies," *arXiv preprint arXiv:2009.03362*, 2020. [Online]. Available: https://arxiv.org/abs/2009.03362

[6] Y. Zhou, X. Li, and W. Lu, "Empirical Study of Financial Crises Based on Topological Data Analysis," *Physica A*, vol. 559, p. 125022, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437120304994

[7] A. Goel, D. Filipović, and P. Pasricha, "Sparse Portfolio Selection via Topological Data Analysis based Clustering," *arXiv preprint arXiv:2401.16920*, 2024. [Online]. Available: https://arxiv.org/abs/2401.16920

[8] A. Hatcher, *Algebraic Topology*, Cambridge University Press, 2002

[9] J. R. Munkres, *Topology, vol. 2*, Prentice Hall, Upper Saddle River, 2000.

[10] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, "Persistence barcodes for shapes," *Int. J. Shape Model.*, vol. 2, no. 5, pp. 99–110, 2004.

Using the BBC Micro:bit in Educational Settings: Recommendations for N. Macedonia

Gorica Miceva
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
gorica.gjorgieva@gmail.com

Mile Jovanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
mile.jovanov@finki.ukim.mk

Emil Stankov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
emil.stankov@finki.ukim.mk

Zivko Atanaskoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
zivko.atanaskoski@finki.ukim.mk

***Abstract*—The integration of the BBC Micro:bit into educational settings has been gaining momentum across various countries due to its potential to foster computational thinking, digital literacy, and hands-on learning. This paper examines the role of the Micro:bit in enhancing STEM education through case studies from Slovakia, Sweden, and the UK. The study explores teaching approaches, technical considerations, student engagement, pedagogical insights, challenges, and cultural contexts, offering valuable insights into the effectiveness of the BBC Micro:bit in different educational environments. Further, based on the findings, we propose a set of recommendations for N. Macedonia.**

***Keywords*—BBC Micro:bit, STEM education, computational thinking, educational technology**

I. Introduction

The BBC Micro:bit is a pocket-sized programmable device designed to introduce young learners to coding, problem-solving, and computational thinking. In addition to its widespread use in formal education, Micro:bit has also fostered an ecosystem of community-driven learning, where extracurricular activities, coding clubs, and online platforms have further empowered students to explore technology outside the classroom. The device's simple yet versatile nature makes it an ideal tool not only for introducing foundational programming concepts but also for encouraging more advanced skills in computational thinking. Furthermore, the Micro:bit's emphasis on hands-on, project-based learning allows students to see the tangible results of their coding efforts, which in turn enhances motivation and engagement. This combination of accessibility, practicality, and creative potential makes Micro:bit an invaluable resource in modern educational settings, facilitating the development of a new generation of problem solvers and innovators.

Since its introduction, Micro:bit has been incorporated into various educational settings worldwide, including in Slovakia, Sweden, and the UK. These countries have integrated Micro:bit into their school curricula to develop key 21st-century skills, including digital literacy, creativity, and critical thinking [1], [2], [3]. This paper presents an analysis of the usage of the device to promote STEM education, engage students, and support teaching practices in these diverse contexts.

The paper proceeds as follows. Section II discusses the related work on the topic. Section III presents a case study on how Micro:bit has been integrated into educational systems in Slovakia, Sweden, and the UK, demonstrating its widespread use in fostering STEM education. The device's integration is analyzed from different perspectives: teaching approaches, technical considerations, student engagement, pedagogical insights, challenges, and cultural context. Section IV provides recommendations for our country – N. Macedonia, based on the experiences of the three countries under consideration. These recommendations emphasize the importance of teacher training, increasing access to devices, and promoting interdisciplinary learning to maximize the impact of the Micro:bit on education. The final section of the paper provides concluding remarks.

II. Related Work

The BBC Micro:bit has gained widespread adoption across the globe, becoming an essential tool for fostering computational thinking and digital literacy in educational settings. Researchers have examined its use in various countries, highlighting the impact on STEM education and its ability to engage students in hands-on learning.

In Germany and France, studies have demonstrated the device's effectiveness in supporting interdisciplinary learning. For instance, studies have discussed how the Micro:bit is used in various countries to teach subjects beyond traditional STEM fields. In Germany, teachers have integrated the device into art and music classes, where students use the Micro:bit to create interactive art installations and music-based projects. This creative use of technology highlights the versatility of the device in bridging different subject areas and fostering innovative thinking among students [1].

In Greece, the BBC Micro:bit has been recognized as a valuable tool for teaching not only programming but also robotics and electronics [3]. Research has emphasized how the device serves as an accessible entry point into more complex STEM fields. Through building simple robotic systems and engaging in hands-on electronics projects, students gain practical skills that can lead to further exploration in engineering and technology.

In Taiwan, studies have shown that the Micro:bit can be used to improve computer literacy, with a particular focus on increasing student motivation and performance in

introductory programming courses [4]. The device's ability to support a range of programming languages, from block-based coding to more advanced text-based languages, has made it an effective tool for engaging students of various skill levels. Taiwanese educators have used the Micro:bit to teach both fundamental and advanced programming concepts in a way that is accessible and engaging for young learners.

In Turkiye, research found that the integration of Micro:bit into classrooms helped students enhance their problem-solving skills and foster a deeper understanding of STEM concepts. Through hands-on, project-based learning, Turkish students were able to visualize the outcomes of their programming efforts, which contributed to increased motivation and a better understanding of how technology can solve real-world problems [5].

Additionally, a systematic review by Gursoy and Arslan [6] explored the broader use of the BBC Micro:bit in education. This review highlighted its widespread adoption across various countries and analyzed its role in fostering both technical and creative skills among students. The findings indicated that the device's flexibility and ease of use made it particularly effective in engaging students across diverse educational contexts, especially in terms of improving computational thinking and enhancing STEM education.

Another review [7] also highlighted how Micro:bit has been adopted in the UK for both educational and extracurricular purposes, focusing on the benefits it offers for practical, hands-on learning. The flexibility of the device enables students to explore computational thinking in creative ways, contributing to an enhanced learning experience.

Johnson [8] examined how Micro:bit has been integrated into classroom environments, emphasizing the device's potential for fostering teamwork and problem-solving skills. This study noted that the hands-on nature of Micro:bit activities encourage students to collaborate and work through challenges together, an essential skill for the future workforce.

Furthermore, studies have explored how the Micro:bit has helped improve engagement in computer science education. By introducing students to coding in an accessible and fun way, it was found that the device encouraged greater interest in programming and helped build foundational skills for future coding endeavors [9].

An exploration of the technical advantages of Micro:bit, particularly its simplicity in teaching physical computing, showed that the device is an effective introduction to the world of electronics and robotics for young learners, enabling them to construct projects that incorporate both software and hardware [10].

A case study by Baxter [11], focused on the impact of Micro:bit in primary schools, found that students not only learned valuable coding skills but also gained confidence in using technology to solve real-world problems, thus broadening their understanding of STEM subjects.

Additionally, educational robotics and the integration of Micro:bit into K-12 classrooms has received attention in research, as it highlights the device's role in teaching not just coding, but also physical computing and creative applications. The integration of robotics with Micro:bit can foster students' problem-solving and engineering skills, contributing to the development of interdisciplinary knowledge in STEM subjects [12].

Finally, an examination of the educational implications of the BBC Micro:bit [13] has highlighted how the device has been successfully used to improve both teaching strategies and student learning outcomes in multiple educational systems. It has been emphasized that the device's low cost and ease of use make it an attractive option for schools with limited resources, helping bridge the digital divide and promote equity in education.

III. A Case Study

In this section we compare the use of Micro:bit in Slovakia, Sweden, and the UK, based on teaching approaches, technical considerations, student engagement, pedagogical insight and cultural context.

A. Teaching Approaches

Across the countries in question educators have employed different teaching strategies using the Micro:bit to enhance learning outcomes.

- Slovakia: The Slovakian experience focuses on teaching programming through Scratch and Python in both elementary and secondary schools. Hands-on learning using block-based and text-based languages provides students with an accessible entry point into coding [3].
- Sweden: Swedish educators have focused on computational thinking and iterative design, often using feedback loops in workshops to refine lesson plans and improve teaching practices. This iterative approach helps students develop a deeper understanding of problem-solving and design [6].
- UK: In the UK, Micro:bit is primarily used in physical computing activities for younger students (ages 7-9), encouraging creative projects and teamwork. Students engage in real-world problem solving, fostering creativity and collaboration [7].

B. Technical Considerations

The technical strengths of Micro:bit have made it an attractive option for educators, but several challenges related to curriculum alignment and access to technology have arisen in each country.

- Slovakia: The Micro:bit's affordability and accessibility make it an ideal tool for schools with limited budgets. However, the lack of a unified programming environment has led to inconsistencies in its use across different schools [3].
- Sweden: While the iterative design process provides flexibility, aligning teaching materials with the new curriculum standards has been a challenge. Nevertheless, the device's versatility allowed teachers to adapt their approach effectively [6].
- UK: The UK has made use of Micro:bit's ability to support multiple programming languages like Scratch and Python, although logistical issues, such as unequal access to devices, presented obstacles to widespread use [7].

C. Student Engagement

Micro:bit has proven effective in keeping students engaged through interactive, hands-on learning experiences across the three countries.

- Slovakia: The device has been a powerful tool for motivating students to explore programming and algorithmic thinking. The hands-on approach has increased student interest in technology and coding [3].
- Sweden: In Sweden, students find activities challenging yet enjoyable. The iterative learning process helps keeping them engaged as they navigate problem-solving tasks and design challenges [6].
- UK: The interactive nature of the Micro:bit has been key in sparking curiosity among students, especially through creative projects that combine physical computing with coding [7].

D. Pedagogical Insights

The pedagogical approaches to teaching with the Micro:bit emphasize hands-on, problem-based learning, which fosters creativity, teamwork, and critical thinking.

- Slovakia: In Slovakia, project-based learning is central to the educational approach, with a combination of block-based and text-based programming allowing students to progress at their own pace while learning new skills [3].
- Sweden: Sweden's focus on computational thinking has been integrated into the curriculum, and the iterative design approach has allowed teachers to refine their materials and improve engagement [6].
- UK: The UK has developed three distinct teaching approaches: inspiring, providing, and consuming, which together encourage students to think creatively and critically while solving problems [7].

E. Challenges and Opportunities

While the integration of Micro:bit in educational systems has been largely successful, it has also presented challenges related to access to technology, curriculum alignment, and teacher support.

- Slovakia: Despite challenges in creating a unified national programming environment, the flexibility of the Micro:bit has provided opportunities for innovation in STEM teaching. Schools have been able to experiment with different approaches and adapt the device to suit their specific needs [3].
- Sweden: The challenge of aligning teaching materials with new curriculum standards was addressed through the iterative feedback loop, allowing teachers to refine their methods over time and better meet the needs of students [6].
- UK: Logistical challenges, such as unequal access to devices, continue to be a barrier to full integration. However, the versatility of the Micro:bit in supporting various educational contexts has led to numerous opportunities for creative learning [7].

F. Cultural Context

The educational context in each country has shaped how the Micro:bit is used in schools, with different cultural values influencing the device's integration into curricula.

- Slovakia: The affordable and flexible nature of Micro:bit fits well within the Slovakian educational context, where cost-effective solutions are often prioritized. The device's low cost has enabled widespread use in schools [3].
- Sweden: Sweden's strong focus on adapting to new curriculum standards has allowed the country to incorporate computational thinking into the educational system effectively. The iterative nature of the approach also aligns well with Sweden's educational culture of continuous improvement [6].
- UK: In the UK, the focus on structured physical computing activities and the integration of Micro:bit into broader STEM education has emphasized the development of creativity and critical thinking skills [7].

IV. Recommendations for N. Macedonia

Table I presents a summary of the important role that the BBC Micro:bit has played in enhancing STEM education in the three countries under consideration, with derived recommendations for its use in N. Macedonia. Based on the experiences of Slovakia, Sweden and the UK, we recommend:

- **Focus on Teacher Training**: Ongoing professional development for educators is essential. Macedonia should prioritize providing teachers with the tools and resources to confidently use Micro:bit in the classroom, including access to training materials and workshops.
- **Increase Access to Technology**: Ensuring equitable access to Micro:bit devices in schools is crucial. Partnerships with international organizations and local businesses could help bridge gaps in device availability.
- **Promote Interdisciplinary Learning**: Macedonia can integrate the Micro:bit into a broader STEAM curriculum, encouraging creativity and critical thinking across subjects such as art, music, and science. Interdisciplinary projects could engage students and help them see the practical applications of coding.
- **Foster Collaboration and Innovation**: Encouraging collaborative, project-based learning can help students develop teamwork and problem-solving skills while using the Micro:bit. Schools should create opportunities for students to work together on real-world challenges, helping them apply their knowledge in meaningful ways.

By embracing these recommendations, Macedonia can enhance its efforts to develop digital literacy, computational thinking, and creativity among students, preparing them for the challenges of the future.

V. Conclusion

In this paper we examined the role of the Micro:bit in enhancing STEM education. The BBC Micro:bit has proven to be a valuable educational tool in Slovakia, Sweden, and the UK, offering numerous opportunities for student engagement, creativity, and interdisciplinary learning. While challenges related to curriculum alignment, access to technology, and teacher support persist, the potential for enhancing STEM education through Micro:bit is evident across all these countries. The summarized findings are presented in Table 1,

and additional recommendations are provided. With these recommendations, Macedonia can enhance its efforts to develop digital literacy, computational thinking, and creativity among students, preparing them for the challenges of the future.

ACKNOWLEDGMENT

The research presented in this paper is partly supported by the Faculty of Computer Science and Engineering, at Ss. Cyril and Methodius University in Skopje.

REFERENCES

[1] W. Holmes, The impact of educational technologies in the 21st century. Cambridge University Press, 2019.

[2] S. Papadakis and M. Kalogiannakis, "Micro:bit in education: A review of literature," Int. J. Technol. Educ. Sci., vol. 1, no. 2, pp. 77-92, 2017.

[3] M. Kalogiannakis, S. Papadakis, and E. Tafa, "The integration of Micro:bit into K-12 education: An investigation in Greece," Comput. Educ., vol. 151, p. 103844, 2020.

[4] Y. Wen, H. Wu, and J. Hsu, "The use of BBC Micro:bit in promoting computational thinking: A case study in Taiwan," Educ. Media Int., vol. 57, no. 3, pp. 259-272, 2020.

[5] S. A. Dymond and B. M. Chandler, "Impact of Micro:bit on student learning outcomes in programming," Comput. Educ. J., vol. 45, no. 1, pp. 10-21, 2018.

[6] E. Gursoy and R. Arslan, "A systematic review on the use of BBC Micro:bit in education," Educ. Media Int., vol. 56, no. 1, pp. 1-14, 2019.

[7] C. Davies and K. Tredwell, "Developing coding and STEM education in the UK through Micro:bit," Brit. Educ. J., vol. 49, pp. 22-30, 2019.

[8] R. Johnson, "Teaching coding with Micro:bit," Comput. Sci. Educ., vol. 27, no. 3, pp. 155-173, 2021.

[9] B. Anderson and M. Rose, "Exploring the impact of Micro:bit in primary schools," Interdiscip. J. Comput. Educ., vol. 9, pp. 49-61, 2021.

[10] M. S. Lee, :Assessing the effectiveness of physical computing tools in the classroom," J. Comput. Educ., vol. 1, no. 4, pp. 354-369, 2020.

[11] A. Baxter, "Hands-on learning with the BBC Micro:bit in primary education," J. Educ. Technol. Soc., vol. 25, no. 1, pp. 52-64, 2020.

[12] S. Papadakis and M. Kalogiannakis, "Educational robotics: Integrating Micro:bit into K-12 classrooms," Educ. Technol. Res. Dev., vol. 67, no. 4, pp. 615-630, 2019.

[13] Y. Gülbahar and R. Yılmaz, "A review of the integration of the BBC Micro:bit in educational contexts," J. Educ. Technol. Soc., vol. 23, no. 4, pp. 45-60, 2020.

TABLE I. SUMMARY OF DIFFERENT DIMENSIONS OF THE USE OF BBC MICRO:BIT IN ENHANCING STEM EDUCATION IN SLOVAKIA, SWEDEN, AND UK, WITH DERIVED RECCOMENDATIONS FOR NORTH MACEDONIA

Dimension	Slovakian Experience	Swedish Experience	UK Experience	Recommendations for N. Macedonia
Teaching approaches	Focus on programming with Scratch and Python, with a hands-on approach in both elementary and secondary schools.	Focus on computational thinking and iterative design, with workshops refining teaching strategies.	Physical computing for younger students (ages 7-9), with emphasis on creativity and problem-solving.	- Integrate hands-on learning in programming with both block-based and text-based languages. - Emphasize interdisciplinary STEAM approaches to encourage creativity and critical thinking.
Technical considerations	Micro:bit's affordability and ease of use; however, the lack of a unified programming environment caused inconsistencies.	Flexibility in the iterative design process, but challenges with aligning teaching materials with new curriculum standards.	Multiple programming languages supported (Scratch, Python), though logistical challenges with device access were present.	- Focus on increasing access to Micro:bit devices. - Strengthen technical support and offer a unified programming environment.
Student engagement	High levels of student motivation and interest in programming and problem-solving.	Students enjoyed the challenge of activities, with feedback loops enhancing engagement.	The interactive nature of Micro:bit kept students engaged through creative projects.	- Emphasize creative, hands-on projects that allow students to explore coding and problem-solving. - Create opportunities for real-world applications of coding.
Pedagogical Insights	Project-based learning using a mix of block-based and text-based programming.	Computational thinking integrated into curriculum; iterative teaching processes improved engagement.	Three teaching approaches: inspiring, providing, and consuming, fostering creativity and problem-solving.	- Provide training for teachers to implement project-based and iterative learning approaches. - Develop interdisciplinary lesson plans that integrate coding with subjects like art, music, and science.
Challenges	Lack of a unified programming environment; budget constraints.	Alignment with new curriculum standards; adapting to evolving technology requirements.	Logistical issues with device access in some areas; unequal distribution of resources.	- Improve teacher support with continuous professional development. - Work with international organizations for broader access to devices.
Opportunities	Flexibility of Micro:bit offers innovative opportunities for STEM teaching.	Iterative feedback allowed for continuous improvement in teaching materials and student engagement.	Micro:bit's versatility opened doors for diverse educational applications in both formal and informal contexts.	- Encourage collaboration and innovation through interdisciplinary projects. - Partner with international organizations (like the British Council) to support the digital transformation of education.
Cultural context	Micro:bit's low cost and flexibility fit well within Slovakia's educational framework.	Sweden's focus on adapting to curriculum changes aligned with computational thinking.	Focus on structured physical computing within the broader context of STEM education.	- Leverage international support to enhance North Macedonia's digital education efforts. - Create a cultural shift in schools by integrating coding and creativity into all areas of learning.

Exploring the Educational Potential of Virtual Reality and Mixed Reality: Immersive Learning, Student Engagement, and Knowledge Retention

Mila Dodevska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
mila.dodevska@finki.ukim.mk

Zivko Atanaskoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
zivko.atanaskoski@finki.ukim.mk

Mile Jovanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
mile.jovanov@finki.ukim.mk

Emil Stankov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
emil.stankov@finki.ukim.mk

Petre Lameski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
petre.lameski@finki.ukim.mk

***Abstract*—The integration of Virtual Reality (VR) and Mixed Reality (MR) technologies in education presents new opportunities for immersive and interactive learning. This paper reviews recent applications of VR/MR in educational contexts, emphasizing their impact on student engagement, cognitive development, and knowledge retention. The analysis highlights key benefits such as enhanced motivation, improved practical skills, and effective visualization of abstract content, while also acknowledging limitations including cognitive load and motion sickness. In addition to the literature review, the grounds of an experimental study are presented.**

***Index Terms*—Virtual Reality (VR), Mixed Reality (MR), Immersive Learning, Student Engagement, Knowledge Retention, E-learning, Educational Technology.**

I. Introduction

Quality education is essential for fostering economic growth, reducing inequalities, and empowering individuals with the knowledge and skills needed for sustainable development [1]. The level of education within a community directly influences its social progress, workforce competence, and overall quality of life, reinforcing the idea that investing in education leads to long-term societal and economic benefits [2].

Technology and education have become inseparable. One of the roles of technology in education is to serve as a tool during teaching and learning [1]. The face-to-face (f2f) learning process was replaced by online or mixed learning. Learning media also evolved from paper-based media to digital-based media and cloud-based media. Digital-based media refers to any electronic format used for learning, replacing traditional paper-based materials. Cloud-based media involves storing and accessing educational content online via cloud computing services instead of local storage. Cloud-based media allows students and teachers to collaborate in real time, access resources from anywhere, and save progress automatically without relying on physical storage. AR/VR has become the most crucial technological resource for stimulating unlimited creativity and facilitating knowledge through experience.

Virtual reality (VR) technology has been used to create situated and realistic learning contexts that learners cannot easily access. One of the examples of VR usage is providing access to scenes and environments that are difficult to observe in the real world, such as surgical procedures — which are typically inaccessible to individuals outside the medical profession. VR can also simulate micro or macro phenomena, such as blood cells or the solar system, making abstract or otherwise unobservable concepts more tangible, as mentioned by [3]. As VR is expected to be widely used in classrooms and adopted by students, particularly over the next couple of years [4] [5], it is argued that, if these predictions hold true, VR technology has the potential to transform teaching pedagogy in ways that are currently unimaginable [6].

The use of computer technology can be perceived through an external component or device, best known as a VR helmet or head-mounted display (HMD). A head-mounted display allows users to experience the virtual surroundings thoroughly as long as they wear the helmet [7]. HMDs extend the advantages of traditional VR in terms of visualization and interaction by tracking head movements to display the corresponding 3D scenes in front of one's gaze and allowing natural manipulation in virtual spaces. In particular, [5] states that VR allows users to experience and feel environments where they are not physically present and manipulate objects in the environment.

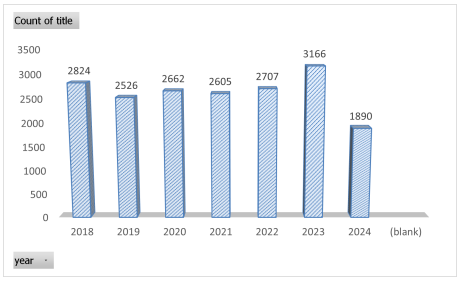


Fig. 1. Number of papers on the topic by each year

For instance, we can grab an object in a VR environment with the help of virtual hands that look like our hands in the real world, move it, lift it, and drop it by moving it to a different point in the virtual environment [7].
In this paper we make a literature review on the topic, and propose further experiment. The paper continues as follows. In the second section we briefly explain our aproach for the literature review. In the following two sections the literature review is presented, and in the fifth section we expose the future steps of our research.

II. Our approach for the literature review

This section describes the systematic process used to identify the most relevant recent contributions published in the literature. The compilation of applicable literature was facilitated through the online academic service Google Scholar, which served as the sole academic database for this research. Publications were searched containing all keywords related to virtual reality applications specifically for educational purposes: "virtual reality in education". Results from Google Scholar were sorted according to relevance, and only freely accessible papers without a paywall that explicitly included all keywords in their titles were selected. Publications considered for review were limited to the period between 2018 and 2025, as earlier contributions were excluded due to the rapid recent advancements in this area.

A spreadsheet was created to systematically organize and evaluate all relevant identified papers, visually shown in Fig. 1, with the focus strictly narrowed down to systematic review papers and experimental studies. A spreadsheet including the key experimental studies that influenced our study are shown in Table I. A visual representation of all of the key literature review studies is omitted due to restrictions in the paper size.

III. Literature Review

Virtual Reality (VR) has gained significant attention in the field of education due to its potential to enhance learning experiences. Several studies have explored the application of VR in various educational settings.

The benefits and potential impact of VR in education have been highlighted in previous research [8]. VR is one of the technologies that could overcome the limitations of space and time by providing numerical simulation and an immersive, interactive environment [9]. It can bring users into an environment that was not generally available because of price, security, or other obstacles [7].

Kong [10] conducted an empirical study on the effects of VR on experiential education, suggesting that more research is needed to determine the effectiveness of such approaches. There is still need to find out whether VR implementation could develop one's skills and how effective VR is compared to other methods of learning and training [7] [11].
There are, in our opinion, relatively few studies that compare the drawbacks of conventional teaching strategies with game-based or immersive learning that makes use of Extended Reality (XR) technologies. Prior research [12] [13] has highlighted how VR technology can complement various learning theories, such as multimedia learning, situated learning, and embodied learning, by simulating the real world and improving learner performance [3].

Wu, Yu, and Gu in [3] focused on the development and student perception of VR specifically for implant surgery, showcasing how VR can be used for specialized training purposes.

Han in [14] reviewed the application of VR in engineering education, highlighting the benefits of VR as an assistive technology in this field. Wei and Yuan [15] summarized the current research status and future trends of immersive VR in education, suggesting potential scenarios for the application of Metaverse and gamification in educational settings.

Sports training has been a general VR application, considering that VR technologies help users to analyze and improve their performance in respective sports branches [7] [16] [17].

Feng, Yuhuan, et al. in [18] focused on the integration of computer VR technology in college physical education, emphasizing the use of Web-based VR technology for teaching and practice. [19] explored the development of health-preserving competence in physical education teachers using VR technologies based on N. Bernstein's theory of movement construction. The study found that VR models were effective tools for enhancing teacher competence.
Putranto, et al. in the study [7] highlight motion capture devices as one of the most commonly used tools to enhance the implementation of VR technology. Motion capture enables the digital recording of human movements, which is particularly useful in sports for tracking and analyzing athletes' actions. For instance, in baseball VR training, motion capture can track a player's swing to estimate the ball's speed and landing position. This technology allows athletes to train flexibly, regardless of location, as long as the necessary equipment is available. Additionally, motion capture is suggested as a valuable tool for improving athletes' skills by providing precise movement analysis during training.

TABLE I
SPREADSHEET SUMMARIZING KEY PAPERS.

Author(s)	Year	Title	Sample Size	Key Outcome	Limitations
Anggara, Risan Putra, et al.	2021	Application of VR and AR in elementary science education	50 students (varied by method)	Hybrid (teacher + VR/AR) had highest improvement (24%) in comprehension	Small sample, early-stage platform
Cieri, Robert L., et al.	2021	VR and AR in morphology and medical education	Multiple case studies	Improved visualization and understanding in anatomy and paleontology	Exploratory case studies, no control group
Remolar, Inmaculada, Cristina Rebollo, and Jon A. Fernandez-Moyano.	2021	Learning history using virtual and augmented reality	50 students	VR/AR improved retention; 40% of students scored perfectly vs. 0% in control group	Navigation issues in VR, limited generalizability
de Moraes Rossetto, Anubis G., et al.	2023	An analysis of the use of augmented reality and virtual reality as educational resources	Not tested in classroom	AR/VR platform for visualizing planets could enhance astronomy education	Tool not yet classroom-tested
Zwolinski, Grzegorz, et al.	2022	Extended reality in education and training: Case studies in management education	Multiple case studies	XR increased engagement, interest, and mood by at least 10%	Limited to short-term feedback, no long-term outcome data
Parong, Jocelyn, and Richard E. Mayer.	2018	Learning science in immersive virtual reality	Lab experiment	Immersive VR improved engagement but not always retention	Context-specific outcomes
Palmisano, Stephen, Robert S. Allison, and Juno Kim.	2020	Cybersickness in head-mounted displays is caused by differences in the user's virtual and physical head pose.	Experimental study	Mismatch in pose tracking causes motion sickness	Limited to sensory conflicts, not learning outcome
Truchly, P., et al.	2018	Virtual reality applications in STEM education	Survey & pilot studies	VR enhances practical skills and concept understanding in STEM	Variation across disciplines
Han, Insook.	2020	Immersive virtual field trips in education: A mixed-methods study on elementary students' presence and perceived learning.	Elementary students (N not given)	Presence and perceived learning improved in immersive field trips	Subjective measures only
Cooper, Grant, et al.	2019	Using virtual reality in the classroom: preservice teachers' perceptions of its use as a teaching and learning tool.	Survey study	Preservice teachers view VR positively as a pedagogical tool	Attitudinal focus, not classroom-tested

In sports, virtual reality is being promoted as a way for athletes to reduce the risk of injury during training. Most sports education and training are originally located in an open room where many obstacles can be encountered [7].
Researchers in the study [7] discuss how various business sectors leverage VR technology to enhance their services. For example, in the healthcare industry, RelieVRx, an immersive multisensory virtual reality therapy, is used at home for managing chronic lower back pain. This prescription-based treatment, approved by the FDA, offers an alternative pain relief method for adults. In the tourism sector, the global travel agency Thomas Cook introduced a VR experience called 'Try Before You Fly,' allowing potential travelers to virtually explore their desired destinations before booking a trip.

Since students should learn from their experience when they are actively involved in solving problems, virtual reality applications and games have a potential to support students' motivation to learn significantly and even to improve their social, collaborative, and cognitive skills [20] [21]. While immersion is defined as the sensitivity level of the sensory features offered by VR systems to users, presence is evaluated as the users' psychological reactions to the virtual environment [7].

A contemporary perspective suggests that immersive learning involves engaging learners in virtual environments that simulate real-world experiences, thereby enhancing cognitive processes essential for learning. This approach leverages technologies such as virtual reality (VR) and augmented reality (AR) to create interactive and realistic scenarios, facilitating deeper understanding and retention of information. Furthermore, recent studies have explored the impact of immersive learning environments on cognitive states. For instance, research has indicated that VR training can enhance the sense of presence, particularly in terms of sensory and realism factors, leading to improved learning efficacy in mechanical assembly tasks [22].

Furthermore, it's essential to highlight the Cognitive-Affective-Social Theory of Learning in Digital Environments (CASTLE), which illustrates how immersive learning experiences are influenced by cognitive, affective, and social factors. Cognitively, learners engage in problem-solving and knowledge retention, while affective factors such as motivation and emotional connection influence engagement. Social interaction is also essential for improving learning outcomes in virtual environments. This includes peer collaboration and instructor guidance [23]. Although the authors consider technologies like AR and VR to be more interactive among participants of the learning process, there are several studies that imply the downsides of these modern technologies usage. Multiple research studies indicate that participants using head-mounted display (HMD)-based immersive learning systems may experience increased cognitive load and motion sickness.
For instance, a study found that complex motor tasks performed in immersive virtual reality (VR) using HMDs led to higher cognitive load and decreased motor performance compared to conventional computer screens [24].

It is worth noting that some studies reported that participants in HMD-based immersive learning also had higher levels of motion sickness and cognitive load than those in non-immersive learning [25]. The immersive nature of HMDs can induce motion sickness, often referred to as cybersickness, due to sensory conflicts between visual inputs and physical motion [25]. These factors are important to consider when implementing HMD-based immersive learning environments.
The integration of game-based learning with Head-Mounted Display (HMD) technology enhances immersion and role-playing opportunities, allowing learners to engage in vivid, realistic scenarios that evoke emotions such as empathy and professional identity [27].
Studies indicate that immersion plays a crucial role in psychological engagement, contributing to a greater sense of

presence, increased motivation, and a more positive attitude towards learning [28]. Additionally, immersive experiences in VR have been linked to improved learning perceptions and subject retention, making them valuable tools for educational and training environments. [29] addressed the ethical and social dimensions of using VR in education, emphasizing the importance of considering socio-ethical challenges associated with the integration of new technologies in educational practices.
Overall, these studies collectively demonstrate the growing interest and potential of VR in enhancing educational experiences and outcomes.

IV. Review of the in class studies

VR applications have been explored in subjects such as science, history, social studies, geography, language learning, and medical training, demonstrating their potential to enhance engagement, retention, and practical skill development. Several studies have investigated VR/AR-based educational tools, highlighting their effectiveness in visualization, simulation, and experiential learning. This section reviews relevant research on VR and AR applications in education, focusing on studies conducted in classes with students - consumers of the introduced technologies.
An important domain where VR applications gradually appear is represented by VR applications suitable for educational deployment [20]. It can be seen that the application of XR in education is mainly related to medical, biology, science, history, social, geography, and language topics. Besides the famous real-life simulators whose usage is widespread (For VR; Nearpod, VR Museum, MelScience, Math VR, Labster; for AR: Google Lens, PokemonGO, Snapchat and Instagram filters, IKEA place), there are several studies on these topics, with an idea similar to our own that are worth mentioning.

Siqueira's study on augmented and virtual reality in astronomy education presents a web-based AR/VR environment designed to help students visualize and interact with graphical representations of planets in the solar system. The authors highlight the potential of this tool for enhancing classroom learning by providing an interactive and immersive experience. However, they note that the system has not yet been tested in an actual classroom setting, leaving its practical educational impact unverified [30]. A 2020 study on virtual planetariums explored the implementation and use of VR technology for teaching astronomy in schools. The developed software aimed to provide an immersive, interactive, and intuitive learning experience, allowing students to navigate through the solar system using VR headsets and auxiliary controls. Users could explore planets and access detailed information about celestial bodies in a simulated environment. Physical questionnaires were given to student users in order to evaluate its efficacy; the results showed that the tool was highly usable and well-received, indicating that it has the potential to improve astronomy instruction [30].

Anggara et al. [31] explored the use of e-learning in science education for third-grade students, focusing on a virtual approach to illustrate real-life events and natural phenomena. Their study involved testing the e-learning platform with a small group of five students and a teacher in an elementary school in Jakarta. The study conducted testing to evaluate the effectiveness of three distinct learning methods for teaching natural sciences to third-grade elementary students: (1) e-learning without teacher explanation, (2) conventional classroom teaching without e-learning tools, and (3) a hybrid approach combining teacher explanation with e-learning methods utilizing Virtual Reality (VR) and Augmented Reality (AR) technologies.
According to the findings, e-learning by itself, without the assistance of an instructor, improved student comprehension by an average of just 8%. Conventional learning methods, relying solely on teacher explanations without digital tools, resulted in an average improvement of 12%. The hybrid learning strategy, which integrates teacher-led instruction with interactive e-learning applications developed using Unity software for Android, had the highest efficacy, resulting in an average enhancement in student comprehension of 24%. The findings suggest that integrating teacher instruction with e-learning methods enhanced by VR and AR technologies significantly increases students' motivation, interest, and overall comprehension of science concepts.

The paper [32] explores four case studies involving different XR technologies applied in education. Each case focuses on a specific learning scenario and uses different AR/VR approaches. The given scenarios were: A projector-based AR system projected guidance on a workspace desk; Fire Safety Training and emergency decision-making; A branching VR scenario where the student played the role of a manager handling an employee's tardiness issue; A mobile-based AR application where students designed product packaging and optimized logistics. Feedback questions for users followed each case study, and student empirical testing revealed that using XR technologies in experiments had a positive impact, increasing mood, engagement, stimulation, and interest by at least 10%.

Another study important to mention is [33], where the authors conducted two experiments to evaluate the effectiveness of a VR and AR-based history learning experience. Although some had minor issues with VR navigation, 25 secondary school students who tested the game's playability, engagement, and usability found it to be very interactive and effective for learning. The second experiment compared AR/VR-based learning with traditional teaching across two student groups (N=50). Results showed that students using VR and AR significantly outperformed those in traditional learning, with 40% achieving perfect scores compared to none in the traditional group. The findings indicate that VR/AR improves enthusiasm, knowledge retention, and inspiration, rendering it a feasible

substitute for traditional approaches. Students had favorable responses to gamification, suggesting that history education can significantly benefit from engaging educational tools.

Cieri et al. in [34] demonstrate the application of VR and AR in morphology, medical training, and education through multiple case studies. VR was used for 3D anatomical visualization, improving pulmonary airflow analysis, cardiac function modeling, and dinosaur track formation studies. Students used interactive virtual reality (VR) modules to learn about dentistry and paleontology, which helped them retain concepts and develop practical skills. The results show that VR/AR offers affordable, immersive, and interactive research and pedagogy options while also greatly improving visualization, analysis, and knowledge acquisition. The study [34] demonstrates how VR and AR, with their user-friendly and collaborative experiences, are revolutionizing morphology research and education. These technologies bridge the gap between the digital and physical worlds, enabling better scientific visualization, collaboration, and learning.

V. Description of the future research steps

VR technologies have great potential to improve conventional teaching strategies by providing dynamic and captivating learning environments, but despite promising results, a direct comparison between traditional learning methods and VR-based learning approaches remains underexplored. Our future work aims to address this question. In the next phase of our research, we plan to conduct an experimental study to systematically compare the effectiveness of conventional teaching methods with immersive VR-based learning. The experiment will involve two groups of students: one using traditional techniques and the other utilizing VR-based environments. Key metrics such as knowledge retention, comprehension, student engagement, and practical skill development will be measured and analyzed.

Additionally, we intend to assess the subjective experience of students participating in VR learning. This evaluation will cover factors such as ease of use, perceived effectiveness, emotional response, motivation, and potential discomfort (e.g., motion sickness). By combining both quantitative performance data and qualitative feedback, we aim to gain a comprehensive understanding of VR's advantages and limitations in educational contexts.

The findings should provide empirical evidence on the effectiveness of immersive technologies, helping educators and policy makers evaluate their advantages over traditional methods and make informed decisions about integrating VR into educational systems.

VI. Conclusion and Future Approaches

The expanding role of virtual reality in education was examined in this research, with an emphasis on information retention, student engagement, and immersive learning. Our analysis of the literature showed how VR technologies have a great potential to improve conventional teaching strategies by providing dynamic and captivating learning environments. While admitting possible drawbacks including higher cognitive load and motion sickness, numerous studies have shown that virtual reality can enhance students' understanding, motivation, and practical abilities across disciplines.

Despite the promising results, a direct comparison between traditional learning methods and VR-based learning approaches remains underexplored. The future steps of our research, explained in the paper, will address these issues. Our aim is to provide empirical evidence on the effectiveness of immersive technologies, helping educators and policymakers evaluate their advantages over traditional methods and make informed decisions on integrating VR into educational systems.

References

[1] Samala, Agariadne Dwinggo, et al. "Global Publication Trends in Augmented Reality and Virtual Reality for Learning: The Last Twenty-One Years." International Journal of Engineering Pedagogy 13.2 (2023).

[2] Berchin, Issa Ibrahim, et al. "The importance of international conferences on sustainable development as higher education institutions' strategies to promote sustainability: A case study in Brazil." Journal of cleaner production 171 (2018): 756-772.

[3] Wu, Bian, Xiaoxue Yu, and Xiaoqing Gu. "Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis." British journal of educational technology 51.6 (2020): 1991-2005.

[4] Zou, Di, Haoran Xie, and Fu Lee Wang. "Future trends and research issues of technology-enhanced language learning: A technological perspective." Knowledge Management & E-Learning 10.4 (2018): 426-440.

[5] Çoban, Murat, Nilüfer Okur Akçay, and İsmail Çelik. "Using Virtual Reality Technologies in STEM Education: ICT Pre-Service Teachers' Perceptions." Knowledge Management & E-Learning 14.3 (2022): 269-285.

[6] Cooper, Grant, et al. "Using virtual reality in the classroom: preservice teachers' perceptions of its use as a teaching and learning tool." Educational Media International 56.1 (2019): 1-13.

[7] Putranto, Jessica Sharon, et al. "Implementation of virtual reality technology for sports education and training: Systematic literature review." Procedia Computer Science 216 (2023): 293-300.

[8] Cvetković, Dragan, ed. Virtual Reality and Its Application in Education. BoD–Books on Demand, 2021.

[9] Feng, Zhenan, et al. "Immersive virtual reality serious games for evacuation training and research: A systematic literature review." Computers & Education 127 (2018): 252-266.

[10] Kong, Lingke "Empirical study on the effects of the application of virtual reality to experiential education on students' learning attitude and learning effectiveness." Revista de Cercetare şi Intervenţie Socială 73 (2021): 288-298.

[11] Stone, Joseph Anthony, et al. "Effectiveness and efficiency of virtual reality designs to enhance athlete development: an ecological dynamics perspective." Movement & Sport Sciences-Science & Motricité 102.4 (2018): 51-60.

[12] Johnson-Glenberg, Mina C. "Immersive VR and education: Embodied design principles that include gesture and hand controls." Frontiers in Robotics and AI 5 (2018): 375272.

[13] Parong, Jocelyn, and Richard E. Mayer. "Learning science in immersive virtual reality." Journal of educational psychology 110.6 (2018): 785.

[14] Han, Yunmeng. "Virtual reality in engineering education." SHS web of conferences. Vol. 157. EDP Sciences, 2023.

[15] Wei, Ziwen, and Man Yuan. "Research on the current situation and future development trend of immersive virtual reality in the field of education." Sustainability 15.9 (2023): 7531.

[16] Cannavò, Alberto, et al. "A movement analysis system based on immersive virtual reality and wearable technology for sport training." Proceedings of the 4th international conference on virtual reality. 2018.

[17] Wang, Bin, and Zheng Zhi Wu. "Research on application of computer virtual reality technology in sports training." Advanced Materials Research 926 (2014): 2694-2697.

[18] Feng, Yuhuan, et al. ”Integration of computer virtual reality technology to college physical education.” Journal of web engineering 21.7 (2022): 2049-2071.
[19] Yevtuch, Mykola, et al. ”Development of the health-preserving competence of a physical education teacher on the basis of N. Bernstein’s theory of movements construction using virtual reality technologies.” Proceedings of the 4th International Workshop on Augmented Reality in Education (AREdu 2021) Kryvyi Rih, Ukraine, May 11, 2021. Vol. 2898. CEUR Workshop Proceedings, 2021.
[20] Truchly, P., et al. ”Virtual reality applications in STEM education.” 2018 16th International Conference on Emerging eLearning Technologies and Applications (ICETA). IEEE, 2018.
[21] Martín-Gutiérrez, Jorge, et al. ”Virtual technologies trends in education.” Eurasia journal of mathematics, science and technology education 13.2 (2017): 469-486.
[22] Lin, Weichao, et al. ”Measuring the sense of presence and learning efficacy in immersive virtual assembly training.” International Journal of Mechanical Engineering Education (2023): 03064190241308659.
[23] Schneider, Sascha, et al. ”The cognitive-affective-social theory of learning in digital environments (CASTLE).” Educational Psychology Review 34.1 (2022): 1-38.
[24] Juliano, Julia M., Nicolas Schweighofer, and Sook-Lei Liew. ”Increased cognitive load in immersive virtual reality during visuomotor adaptation is associated with decreased long-term retention and context transfer.” Journal of NeuroEngineering and Rehabilitation 19.1 (2022): 106.
[25] Makransky, Guido, Thomas S. Terkildsen, and Richard E. Mayer. ”Adding immersive virtual reality to a science lab simulation causes more presence but less learning.” Learning and instruction 60 (2019): 225-236.
[26] Palmisano, Stephen, Robert S. Allison, and Juno Kim. ”Cybersickness in head-mounted displays is caused by differences in the user’s virtual and physical head pose.” Frontiers in Virtual Reality 1 (2020): 587698.
[27] Feng, Zhenan, et al. ”Immersive virtual reality serious games for evacuation training and research: A systematic literature review.” Computers & Education 127 (2018): 252-266.
[28] Han, Insook. ”Immersive virtual field trips in education: A mixed-methods study on elementary students’ presence and perceived learning.” British Journal of Educational Technology 51.2 (2020): 420-435.
[29] Prokopenko, O., and A. Sapinski. ”Using Virtual Reality in Education: Ethical and Social Dimensions.” E-Learning Innovations Journal 2.1 (2024): 41-62.
[30] de Moraes Rossetto, Anubis G., et al. ”An analysis of the use of augmented reality and virtual reality as educational resources.” Computer Applications in Engineering Education 31.6 (2023): 1761-1775.
[31] Anggara, Risan Putra, et al. ”Application of electronic learning by utilizing virtual reality (VR) and augmented reality (AR) methods in natural sciences subjects (IPA) in elementary school students grade 3.” JTP-Jurnal Teknologi Pendidikan 23.1 (2021): 58-69.
[32] Zwoliński, Grzegorz, et al. ”Extended reality in education and training: Case studies in management education.” Electronics 11.3 (2022): 336.
[33] Remolar, Inmaculada, Cristina Rebollo, and Jon A. Fernández-Moyano. ”Learning history using virtual and augmented reality.” Computers 10.11 (2021): 146.
[34] Cieri, Robert L., et al. ”Virtual and augmented reality: New tools for visualizing, analyzing, and communicating complex morphology.” Journal of morphology 282.12 (2021): 1785-1800.

Confidence in the Mirror, Doubt in the Crowd: *Navigating Trust in the Age of AI*

Tanja Pavleska
Laboratory for Open Systems and Networks
Jozef Stefan Institute
Ljubljana, Slovenia
atanja@e5.ijs.si

Vladimir Trajkovik
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of N. Macedonia
vladimir.trajkovik@finki.ukim.mk

***Abstract*—The rapid integration of generative Artificial Intelligence into everyday tasks has brought not only shifts in productivity and creativity but also fundamental changes in how people perceive trust, both in themselves and in others. This study distills key findings from a broader survey-based investigation into public trust surrounding generative Artificial Intelligence tools like ChatGPT. While users report high confidence in their own AI-assisted outputs, their trust in outputs created by others significantly differs. Furthermore, there is a prevalent assumption that high-quality work by others likely involved Artificial Intelligence assistance, often devaluing the human effort behind it. This paper analyzes this asymmetric trust dynamic, exploring its roots, societal implications, and the psychological tension between self-confidence and social skepticism in a technologically mediated landscape. These findings signal a need for trust-aware design, transparency in Artificial Intelligence usage, and strategies to repair and preserve interpersonal trust in Artificial Intelligence-augmented communities.**

Keywords—trust, self-trust, generative AI, distrust, ethics

I. Introduction

Generative Artificial Intelligence (AI) has emerged as a transformative force across diverse domains, from writing and design to programming and education. Its ability to automate and augment human tasks has been embraced by millions of users globally, showcasing an extensive range of use cases, including software development, creative writing, composing, and customer service However, this widespread adoption has also introduced a less visible but equally profound shift: a reconfiguration of trust [1]. Notably, the rise in usage has also sparked concerns about the role of traditional human endeavors and the difficulty in distinguishing between human and AI authorship in creative, academic, and educational contexts [2], [3]. While significant advancements have been observed in service provision, education, and scientific analysis, there are also profound and concerning impacts in areas such as democracy, inequality, security, safety, and social cohesion.

The accelerated evolution of generative AI poses challenges for both developers and the users, as they struggle to keep pace with the latest versions, upgrades and novelties. Some of the users feel overwhelmed by these changes, while others embrace the AI revolution. Some are reluctant to even experience the new AI tools, while others have already integrated them in their daily routines. This discrepancy in perceptions and the resulting adoption behaviors create and widen the gap between adopters and non-adopters, as well as among different types of adopters. This disparity affects not only trust in AI-generated results but also interpersonal trust among users of these tools, potentially leading to social inequality and discrimination. Currently, research primarily scrutinizes trust in AI solutions. However, another dimension of trust, inter-human trust, is being neglected. This form of trust, described as "the lubricant of social interactions" [4] and "the glue that keeps societies from falling apart" [5] remains crucial regardless of whether AI becomes more restricted. A comprehensive examination and analysis are needed to understand the positive and negative social consequences, emerging trends, and areas for improvement in generative models

This paper focuses on the discrepancy between self-trust and interpersonal trust in the context of generative AI, using empirical data to highlight the implications of this divide.

II. Related work

The extensive social literature on trust and its effects on community structure spans decades [4], [6], [7], [8], [9]. Across all disciplines that study it, trust has been recognized as the critical “lubricant” for social interactions in both traditional and computational networks [4], [5].

The body of research on trust in generative AI, particularly focusing on ChatGPT, covers a wide range of domains and provides critical insights into how these technologies are perceived and the implications for their use. Their impact on exacerbating cognitive biases, privacy issues and misinformation spreading has been recognized as major contributors to public distrust in AI [10]. On the other hand, the potential to enhance productivity in various industries, including banking, hospitality, and IT has also been acknowledged.

Recognizing the criticality of ethical and quality issues that are part of the public discourse, Kim et al. explored the impact of these issues in the context of ChatGPT's assistance in travel decisions [11]. Their study revealed that poor-quality responses and unethical aspects significantly reduced travelers' acceptance, satisfaction, and perceived trustworthiness of ChatGPT's recommendations in general, highlighting the sensitivity of users to the affective properties of communication when evaluating AI-generated content. Expanding on this

theme, [12] investigated how user motivations influence perceptions of creepiness[1] and trust in ChatGPT. They found that personalization reduces creepiness, while task efficiency and social interaction increase it. Furthermore, trust is positively impacted by both task efficiency and personalization, which in turn, enhance users' intention to continue using AI tools. This study underscores the importance of understanding user motivations to foster trust and reduce discomfort with AI technologies. In that context, [13] reveals the importance of careful design of generative AI interfaces for enhancing trustworthiness. By comparing trust in ChatGPT-generated content with that from Google Search and Wikipedia, the authors showed that users do not universally trust ChatGPT as much as other sources, with perceived gatekeeping and information completeness being crucial factors in trust formation. Driven by a similar idea, but arriving to somewhat different findings, Huschens et al. investigated the perceived credibility of human versus AI-generated content [14]. Their results indicate that users perceive AI-generated content as similarly credible, competent, and trustworthy as human-generated content, although AI-generated content was viewed as clearer and more engaging. Thus, they too suggest that presentation and engagement are key factors in the acceptance of AI-generated content.

All of these studies provide a nuanced understanding of the trust dynamics and ethical considerations for generative AI, and underscore the importance of quality, ethical standards, and user motivations in fostering trust and maximizing the positive impacts of AI technologies. However, they mainly investigate the human-AI trust, offering little insight into the impact of AI on inter-human trust, and vice versa.

Some efforts in that directions were made in the context of examining the implications of ChatGPT on academic integrity [15], with calls for reconsideration of the definition and the process of acquiring knowledge. Further analyzing the social impact of generative AI, another study emphasize the importance of addressing inequality, bias, discrimination, and stereotypes associated with generative models, and the critical examination of the social factors contributing to these issues, beyond just technical fixes [16]. Finally, recognizing the broad impact of AI on society and the broader accountability gaps in highly autonomous complex systems, [17] propose a new digital right, the 'right to reparation', reflecting the need for robust ethical frameworks as AI technologies become more complex and integrated into society.

While some of the studies reflect on the importance of the broader trust context than just the human-AI relation, there is very poor analysis of that context, especially empirical ones. In the following sections, we will present a wider view of this relationship, making an effort to obtain empirical data on and analyze the trusting sentiments of people regarding AI and its users.

III. Methodology

A. Survey design

In order to investigate the human-to-human trust in the presence of generative AI, we developed a Survey integrating elements of the Likert Scale Methodology [31] and Cross-Sectional Survey Design [32]. We used open questions to investigate contextual traits, and various choice-based questions to provide quantitative and qualitative analysis of the trusting aspects. This allows for extracting both descriptive and measurable determinants of trust.

The survey was distributed through social networks, mail channels, as well as the channels and networks of friends, colleagues and other relevant contacts. The data collection period was 2.4 – 16.4.2024. A total of 202 people provided their responses to the survey within the data gathering period. Data from the survey is available upon request.

Ethical consideration: No direct interviews or personal encounters were conducted for the purpose of data collection in this study. Additionally, no personal data were obtained or processed in any form. The findings are based exclusively on the cumulative results derived from statistical analysis of the survey data. Each participant was fully informed about the purpose of the survey and the associated data processing practices. Consent for participation was obtained from all respondents prior to their engagement in the survey, ensuring compliance with ethical standards.

B. Demographics and context

The geographical spread and profile of the respondents is of general type. The majority are in the range of 26-60 years, more precisely: 44% in the range of 26-40, 45% in the range of 41-60, 6% between 15-25 and 5% above 60. Almost all of them are technically savvy, self-assessing their technical savviness as: 44.5% - Excellent, 33% - Very good, 18.5% - Good, and 4% - Sufficient. Their experience in using generative AI solutions is also at a high level, with only 3% not being familiar with generative and OpenAI, 16.4% who have not used any solution, equal 16.4% who just use it for "playing around", and high 57.2% that have been using it on a more regular basis.

Regarding the type of tool and the purpose of usage, the survey responses indicate that responders widely use AI tool for a variety of purposes. Many use ChatGPT for professional tasks such as preparing presentations, conducting research, writing emails, debugging code, generating initial code bases, and project planning. Furthermore, tools like ChatGPT and DALLE are used for automation, data analysis, writing project grants, academic tasks, and policy suggestions. ChatGPT is also utilized for educational purposes, including studying and learning, for tasks such as generating text summaries, improving written text, and conducting qualitative analysis. Both ChatGPT and Midjourney are popular for creative writing, generating promotional images, and creating content for social media.

[1] "Creepiness" is defined as a feeling of discomfort that is experienced when a user is uncertain how to interact with the AI tool.

While several users reported interacting with ChatGPT and other tools out of curiosity, experimenting with their capabilities for various tasks, the majority have also been testing AI tools for potential applications in professional work and personal projects. Some specific use cases include: using ChatGPT for writing film synopses, sermon preparation, creating advertising texts, and answering domain-specific questions. Other tools like Whisper and GPT-4 are used for coding assistance, generating images, and discovering new applications in research.

IV. RESULTS

The following main questions were the subject of the trust analysis in this paper:

- Q1: How much do you trust AI tools in assisting you in (some of) your tasks?
- Q2: How much do you trust the results (artwork, literature or academic writing, code, etc.) of others that were assisted by AI?
- Q3: How much do you trust others' good results (artwork, literature or academic writing, code, etc.) to have been produced without the assistance of AI?
- Q4: Do you think AI tools should be available to everyone?

A. *Trust in the self vs. Trust in others*

To visualize the discrepancy in the results obtained from the analysis of the answers to Q1 and Q2, we comparatively analyze both of the responses through the same figure - Fig. 1. It compares the levels of trust respondents exhibit toward their own AI-assisted solutions and the solutions of others.

The results reveal a distinct pattern in how trust is distributed across these two categories. Respondents show greater trust in their own AI-assisted solutions, with trust levels peaking significantly at 3, where over 90 respondents report high confidence in their own results. At higher trust levels, such as 4, trust in one's own solutions remains high, gradually diminishing at the extreme levels. In contrast, trust in others' solutions is distributed across mid-range levels, with a significant peak at 3 (but lower than for self-trust), where approximately 75 respondents express moderate trust. However, trust in others' solutions declines sharply at higher levels compared to trust in personal results.

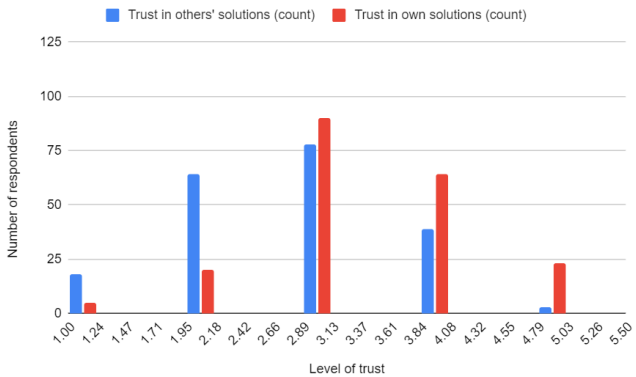


Fig. 1. Comparison of the levels of trust for Q1 (red bars) and Q2 (blue bars)

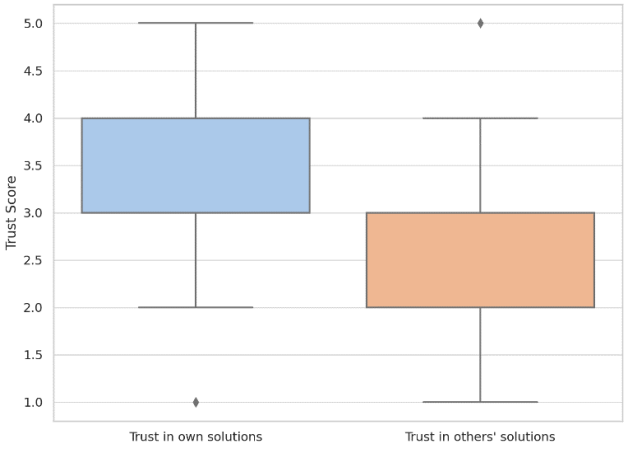


Fig. 2. Difference in opinions for trust in own vs. trust in others' solutions

These trends reflect a strong self-serving bias, as respondents tend to trust their own AI-assisted work more than that of others. The average trust in one's own solutions is centered around higher levels, indicating a greater confidence in personal outputs, while trust in others' solutions is concentrated at lower levels, suggesting skepticism or uncertainty about external work. The distribution of trust in others' solutions also exhibits greater variability, with respondents spreading their trust more evenly across a broader range of levels. The trust levels for both categories converge at 3, but beyond this point, trust in one's own solutions consistently outpaces trust in others' results.

To statistically examine the difference in trust between participants' perceptions of their own use of AI tools and their trust in others' AI-assisted outputs, a paired samples t-test was conducted, allowing a direct comparison of the two conditions. The results revealed a statistically significant difference, with a t-statistic of 8.38 and a p-value of 9.17×10^{-15}. This is further visualized in the simple boxplot on Fig. 2. The extremely low p-value confirms that the observed difference is not due to random chance.

This pattern highlights a notable disparity in how trust is allocated, emphasizing the impact of self-confidence and skepticism toward external work. The results suggest that while AI may enhance users' confidence in their own outputs, doubts persist about the reliability or authenticity of AI-influenced work by others. This dynamic emphasizes the importance of fostering shared standards and improving transparency in AI-generated content to address these trust disparities and build a more cohesive understanding of AI's role in collaborative environments.

B. *Implications of interpersonal distrust*

When asked whether they trust that others' good results have been produced without the assistance of AI, respondents have expressed notably high distrust, as visualized on Fig. 3.

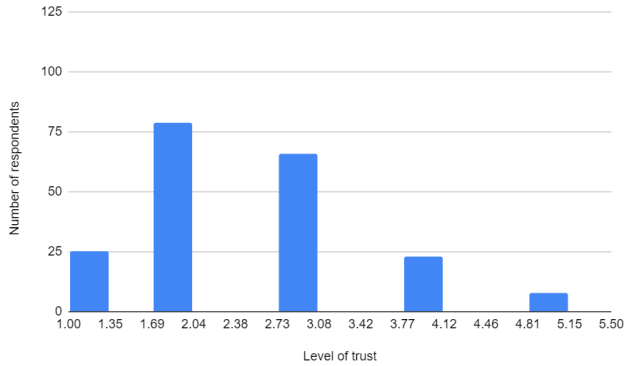


Fig. 3. Level of trust for Q3: "How much do you trust others' good results to have been produced without the assistance of AI?"

The results emphasize the prevailing doubt among respondents about whether others' results/solutions have been obtained with the use of AI. The majority of respondents fall within the low trust part of the distribution, with trust levels peaking around 2 and 3.

These levels represent the highest frequencies of respondents, suggesting that most participants express doubt and distrust in the non-AI origin of others' work. Higher trust levels, such as 4 and 5 are notably less frequent, indicating that strong confidence in non-AI origins is rare among participants. This skepticism may stem from the growing integration of AI tools in various domains, making it increasingly difficult for respondents to trust that others' outputs were created without AI assistance. The low levels of trust also have a smaller but significant portion of respondents, representing those who are highly skeptical about the non-AI origin of solutions.

Quantitatively, the results suggest a central tendency clustered around mid-to-low trust levels, with respondents showing hesitation or doubt about the non-use of AI by others. The variance is moderate, with responses spread across a range of trust levels but concentrated in the lower-to-moderate range. This indicates a broad but skeptical perception of non-AI authenticity. This makes the overall picture of trusting inclinations somewhat alarming, as it demonstrates a state of default low (if not absent) interpersonal trust, and one that is significantly lower than the reported self-trust.

Considering the rapid integration of AI tools across many domains and their widespread use by ordinary users, the presented results from the trust-related analysis appear worrisome. On the one hand, they indicate an increasing reliance on generative AI as facilitators and personal assistants, whereas on the other hand, they denote a significant shift towards default distrust in the actions and intentions of others.

This interpersonal skepticism has far-reaching implications. In creative and academic fields, the assumption that quality equates to AI usage can devalue genuine human effort. In collaborative environments, such as workplaces or classrooms, it may inhibit healthy teamwork and create suspicion among peers. At a societal level, widespread distrust may foster division, inequality, and resistance to technological integration.

Moreover, the tendency to assume AI use by others—even without evidence—risks becoming a self-fulfilling prophecy: if everyone believes others are using AI and distrusts them for it, the environment becomes one of guardedness and opacity. Such a climate stifles innovation and collaboration, which are essential for thriving in an AI-augmented world.

C. Cross-demographic trust analysis

To better understand how demographic factors may influence perceptions of trust in AI, a series of one-way ANOVA tests were conducted. The analysis focused on three dimensions of trust: individuals' trust in their own use of generative AI tools, their trust in others' AI-assisted outputs, and their belief that others' high-quality results were produced without the use of AI. These were examined in relation to two key demographic variables: age group and self-reported technical savviness.

Trust in one's own AI-assisted work was found to vary significantly with technical expertise, indicated by a statistically significant F statistic of 3.084, and a p-value of 0.028. Thus, participants who rated themselves as more technically savvy expressed higher levels of confidence in using AI tools. This is perhaps expected, as familiarity with technology may breed trust in one's ability to use it effectively. Age, however, was not a significant factor in shaping self-trust ($F = 1.263$, $p = 0.288$).

In contrast, the results for trust in others' AI-assisted outputs were strikingly uniform. No statistically significant differences were observed across age groups ($F = 0.916$, $p = 0.434$) or levels of technical savviness ($F = 0.070$, $p = 0.976$). Similarly, regarding Q3, responses showed no meaningful variation based on age ($F = 1.122$, $p = 0.341$) or technical experience ($F = 1.080$, $p = 0.359$). These findings suggest that interpersonal skepticism is broadly distributed and not significantly influenced by demographic factors, pointing to a deeper systemic issue.

D. AI democratization

Finally, we inquired on the respondent's stance on the availability of AI tools and the people's right to access them. Fig. 4 shows the results of this analysis.

A significant majority (73.6%) believe that AI tools should be accessible to everyone, reflecting a strong belief in the democratization of technology and equal access to technological advancements. The sentiment aligns with principles of equality and the idea that technology should benefit all of society, and not just a select few. However, a notable portion (21.3%) of respondents think that AI tools should not be available to everyone, which mainly stems from concerns about misuse, ethical considerations, and the potential negative impacts on various aspects of society, such as job displacement and data privacy (informed by the respondent's comments to the question).

A smaller group (5.1%) believes that AI tools should be available under certain conditions, offering a nuanced view that supports accessibility, but with safeguards to mitigate potential risks and ensure responsible use.

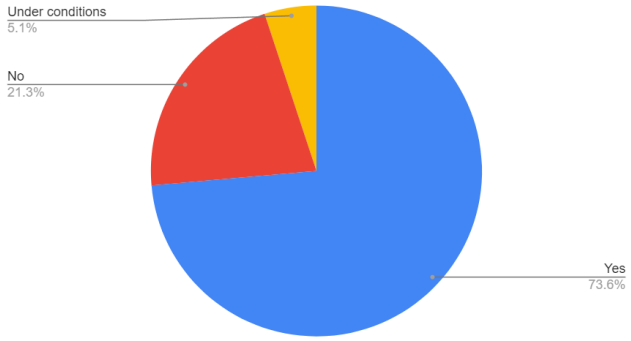


Fig. 4. Attitude towards availability of AI for everyone (Q4)

The rich knowledge base built from the long comment as a response to the open question inquiring for further explanation of the answer to Q4 can be systematized into several key topics and issues:

- *Democratic values and equality:* Several comments emphasize the importance of equal access to AI tools, drawing parallels to democratic principles and the early days of the internet. The respondents' belief is that restricting access could create or exacerbate inequalities and stifle innovation and progress.
- *Dual AI nature:* Some comments recognize both the AI's potential benefits and the risks, which was also reflected in the analysis above, showing a balanced perspective that acknowledges the transformative power of AI while cautioning against unchecked use.
- *Innovation and productivity:* Many respondents see AI as a tool that can significantly enhance productivity and innovation across various fields. There is a strong belief that making AI widely available will lead to widespread benefits, including time savings, improved efficiency, and new opportunities for creativity.
- *Corporate control and ethical use:* Several comments express concern about AI being controlled by large corporations, which could contradict the initial promise of open access. There is a fear that corporate monopolies could lead to misuse or limit the benefits to a broader audience. Ethical use and the potential for AI to exacerbate existing inequalities are also recurring themes across many comments.
- *Education and regulation*: Some respondents point out that while AI should be accessible, there should be education on its use and regulation to prevent misuse. This perspective advocates for a structured approach to AI deployment, ensuring users are knowledgeable and the technology is used ethically.
- *Impact on creativity and human skills*: Among the greatest concerns raised is the one about AI potentially diminishing human creativity and skills due to over-reliance on AI, particularly among younger generations. Thus, there is a call for balance in order to ensure that AI is used as a tool to complement rather than replace human abilities.
- *Open source and transparency*: The advocacy for open-source AI and transparency in its development and deployment reflects a desire for accountability and trust in AI technologies, with open access being pointed out as a way to foster innovation and allow for community-driven improvements and safeguards.
- *Risk of misuse*: Finally, a recurring theme is the potential for AI to be used for harmful purposes, whether by individuals or entities, recognizing the need for proactive risk management and the establishment of ethical guidelines to prevent abuse.

V. Discussion

A. Rebuilding trust in the age of AI

Rebuilding trust in the context of widespread generative AI use requires a multi-layered, human-centered approach that addresses both technological design and social behavior. The discrepancies in how individuals trust themselves versus how they trust others with AI suggest a deeper need to cultivate clarity, empathy, and shared norms around these tools.

A critical starting point is transparency. As generative AI becomes further embedded in everyday workflows, there must be clear and accessible ways to indicate when and how these tools have been used. Whether through embedded markers, voluntary disclosures, or platform-specific policies, transparency can help demystify AI involvement and reduce the suspicion that often clouds interpersonal interactions. When people know the extent to which AI contributed to a result, they are more likely to accept it, evaluate it fairly and engage openly with the creator.

In tandem with transparency, public education and awareness are essential. Mistrust often stems from misunderstanding, and when users lack a realistic picture of what generative AI can and cannot do, their judgments become skewed. Providing accurate, approachable information about how these tools work, their limitations, and their appropriate applications can empower individuals to make informed assessments. This, in turn, fosters a more balanced perspective that tempers fear with understanding and replaces suspicion with discernment.

Beyond education, the social fabric in which AI operates must evolve as well. Building interpersonal trust requires the cultivation of shared norms and expectations. Open conversations about AI use, whether in classrooms, workplaces, or online communities, can help normalize responsible usage while discouraging hidden reliance. As people gain comfort discussing how they use AI and learn to interpret others' use in good faith, a more trusting environment can emerge.

Ultimately, restoring trust in the age of AI is not just about fixing a system but about reimagining our relationships, with technology and with each other. By embracing transparency, fostering understanding, designing for trust, and encouraging honest dialogue, we can begin to rebuild an ecosystem where

self-trust and trust in others not only coexist but reinforce one another. In doing so, we can ensure that AI augments human potential without eroding the social bonds that sustain it.

B. Actionable design recommendations

The responsibility for trust-building also lies at the side of developers. To bridge the interpersonal trust gap identified in this study, designers, developers, and platform providers must rethink how AI is introduced and integrated into user-facing tools. Based on the data analysis and behavioral insights from the survey, some actionable design-specific recommendations can also be made. First, generative AI tools should support transparent disclosure of AI involvement, allowing users to indicate where and how AI contributed to their work. This nuanced transparency, such as attributing specific sections to AI or human and AI, can counteract assumptions that diminish human effort.

Second, incorporating simple explainability features into AI interfaces can build confidence in both users and observers. When systems (even if only briefly) clarify how outputs were generated, it reduces ambiguity and fosters credibility, especially in collaborative or evaluative settings. Ethical nudges, such as prompts encouraging users to disclose AI use prior to sharing content, can help normalize openness without mandating it. Paired with educational onboarding that includes both functional and ethical guidance, these measures can reduce misconceptions and ease discomfort with AI technologies.

Finally, platforms should support light community-based trust mechanisms, such as tools to verify or contextualize AI use. In organizational settings, transparent AI usage dashboards could help teams align on ethical practices and foster mutual accountability. Collectively, these strategies can promote a culture of trust, clarity, and responsible innovation in AI-augmented environments.

VI. Conclusion

The rise of generative AI is not just a technical revolution; it is a social one. As people learn to rely on AI tools, they must also navigate shifting trust dynamics with their peers. This study highlights the critical asymmetry between self-trust and trust in others, showing that while users are confident in their own AI-assisted outcomes, they are skeptical of others'. Addressing this imbalance is essential for cultivating collaborative, equitable, and ethical AI ecosystems. By fostering transparency, ethical design, and shared understanding, we can ensure that generative AI enhances not only our capabilities but also our connections with one another.

References

[1] R. Gozalo-Brizuela and E. C. Garrido-Merchán, "A survey of Generative AI Applications," Jun. 14, 2023, *arXiv*: arXiv:2306.02781. doi: 10.48550/arXiv.2306.02781.

[2] H. Desaire, A. E. Chua, M. Isom, R. Jarosova, and D. Hua, "Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools," *Cell Reports Physical Science*, vol. 4, no. 6, p. 101426, Jun. 2023, doi: 10.1016/j.xcrp.2023.101426.

[3] T. Berber Sardinha, "AI-generated vs human-authored texts: A multidimensional comparison," *Applied Corpus Linguistics*, vol. 4, no. 1, p. 100083, Apr. 2024, doi: 10.1016/j.acorp.2023.100083.

[4] C. Castelfranchi and R. Falcone, *Trust Theory: A Socio-Cognitive and Computational Model*. John Wiley and Sons, 2010.

[5] F. Fukuyama, *Trust: The Social Virtues and The Creation of Prosperity*, 1st Free Press Pbk. Ed edition. New York: Free Press, 1996.

[6] D. Cartwright and F. Harary, "Structural balance: a generalization of Heider's theory," *Psychological Review*, vol. 63, no. 5, pp. 277–293, 1956, doi: 10.1037/h0046049.

[7] F. Heider, *The Psychology of Interpersonal Relations*. Routledge, 1982.

[8] John Conlisk, "Why Bounded Rationality?," *Journal of Economic Literature*, vol. 34, no. 2, pp. 669–700, 1996.

[9] J. H. Fowler and N. A. Christakis, "Cooperative behavior cascades in human social networks," *Proceedings of the National Academy of Sciences*, vol. 107, no. 12, pp. 5334–5338, Mar. 2010, doi: 10.1073/pnas.0913149107.

[10] Y. K. Dwivedi *et al.*, "Opinion Paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102642, Aug. 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[11] J. H. Kim, J. Kim, C. Kim, and S. (Sam) Kim, "Do you trust ChatGPTs? Effects of the ethical and quality issues of generative AI on travel decisions," *Journal of Travel & Tourism Marketing*, vol. 40, no. 9, pp. 779–801, Nov. 2023, doi: 10.1080/10548408.2023.2293006.

[12] T. Hyun Baek and M. Kim, "Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence," *Telematics and Informatics*, vol. 83, p. 102030, Sep. 2023, doi: 10.1016/j.tele.2023.102030.

[13] Y. Jung, C. Chen, E. Jang, and S. S. Sundar, "Do We Trust ChatGPT as much as Google Search and Wikipedia?," in *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, in CHI EA '24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–9. doi: 10.1145/3613905.3650862.

[14] M. Huschens, M. Briesch, D. Sobania, and F. Rothlauf, "Do You Trust ChatGPT? -- Perceived Credibility of Human and AI-Generated Content," Sep. 05, 2023, *arXiv*: arXiv:2309.02524. doi: 10.48550/arXiv.2309.02524.

[15] D. O. Eke, "ChatGPT and the rise of generative AI: Threat to academic integrity?," *Journal of Responsible Technology*, vol. 13, p. 100060, Apr. 2023, doi: 10.1016/j.jrt.2023.100060.

[16] M. T. Baldassarre, D. Caivano, B. Fernandez Nieto, D. Gigante, and A. Ragone, "The Social Impact of Generative AI: An Analysis on ChatGPT," in *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, in GoodIT '23. New York, NY, USA: Association for Computing Machinery, Sep. 2023, pp. 363–373. doi: 10.1145/3582515.3609555.

[17] K. Sharma, F. D. Schoorman, and G. A. Ballinger, "How Can It Be Made Right Again? A Review of Trust Repair Research," *Journal of Management*, vol. 49, no. 1, pp. 363–399, Jan. 2023, doi: 10.1177/01492063221089897.





"Ss. Cyril and Methodius" University in Skopje
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

