



INTERNATIONAL SCIENTIFIC CONFERENCE ON INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND DATA SCIENCE

Sinteza
SINGIDUNUM UNIVERSITY INTERNATIONAL SCIENTIFIC CONFERENCE

Novi Sad, 2025

sinteza.singidunum.ac.rs



Sinteza

SINGIDUNUM UNIVERSITY INTERNATIONAL SCIENTIFIC CONFERENCE

BOOK OF PROCEEDINGS

INTERNATIONAL SCIENTIFIC CONFERENCE ON INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND DATA SCIENCE



Publishing of Conference Proceedings of the International Scientific Conference on Information Technology, Computer Science, and Data Science- Sinteza 2025 has been supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia.

Novi Sad
May 9, 2025.
sinteza.singidunum.ac.rs

SCIENTIFIC COMMITTEE

- Milovan Stanišić, Singidunum University, Serbia
- Endre Pap, Singidunum University, Serbia
- Aleksandar Jevremović, Singidunum University, Serbia
- Andreja Stojić, Institute of Physics in Belgrade, Serbia
- Bratislav Milovanović, Singidunum University, Serbia
- Dragana Četojević-Simšin, Singidunum University, Serbia
- Dragan Cvetković, Singidunum University, Serbia
- Đorđe Obradović, Singidunum University, Serbia
- Eva Tuba, Singidunum University, Serbia
- Goran Šimić, Military Academy, University of Defence, Serbia
- Goranka Knežević, Singidunum University, Serbia
- Gordana Dobrijević, Singidunum University, Serbia
- Gordana Jovanović, Institute of Physics in Belgrade, Serbia
- Jelena Filipović, Faculty of Philology, University of Belgrade, Serbia
- Miloš Antonijević, Singidunum University, Serbia
- Marina Marjanović Jakovljević, Singidunum University, Serbia
- Marko Tanasković, Singidunum University, Serbia
- Marko Šarac, Singidunum University, Serbia
- Milan Tair, Singidunum University, Serbia
- Milan Tuba, Singidunum University, Serbia
- Miljan Vučetić, Vlatacom Institute, Serbia
- Miloš Stojmenović, Singidunum University, Serbia
- Miodrag Živković, Singidunum University, Serbia
- Mirjana Perišić, Institute of Physics in Belgrade, Serbia
- Miroslav Popović, Singidunum University, Serbia
- Mladen Veinović, Singidunum University, Serbia
- Mladan Jovanović, Singidunum University, Serbia
- Nebojša Bačanić Džakula, Singidunum University, Serbia
- Nemanja Stanišić, Singidunum University, Serbia
- Nina Dragičević, Singidunum University, Serbia
- Petar Spalević, Faculty of Technical Sciences in Kosovska Mitrovica, University of Priština, Serbia
- Sanja Filipović, Singidunum University, Serbia
- Saša Adamović, Singidunum University, Serbia
- Tijana Radojević, Singidunum University, Serbia
- Valentina Gavranović, Singidunum University, Serbia
- Zora Konjović, Singidunum University, Serbia
- Živko Bojović, Singidunum University, Serbia
- Alexandru Nedelea, Stefan cel Mare University of Suceava, Romania
- Aurora Pedro Bueno, Department of Applied Economics, University of Valencia, Spain
- Deasún Ó Conchúir, Scatterwork GmbH, Ireland
- Diego Andina De la Fuente, Technical University of Madrid, Spain
- Dragana Vilić, Ekonomski fakultet, Univerzitet u Banja Luci
- Dušan Ličina, École polytechnique fédérale de Lausanne (EPFL), Switzerland
- Duško Lukač, Rheinische Fachhochschule Köln – University of Applied Sciences, Germany
- Egona Lavendelis, Riga Technical University, Latvia
- Gordana Pesaković, Argosy University, USA
- Hong Qi, Dalian University of Technology, China
- Ivan Bajić, Simon Fraser University, Canada
- Ivan Chorbev, Faculty of Computer Science and Engineering, Skopje, North Macedonia
- Ina Bikuvienė, PhD Kauno kolegija- University of Applied Sciences, Kaunas (Lithuania)
- Jovica V. Milanović, University of Manchester, United Kingdom
- Li Liwen, Beijing Foreign Studies University, Beijing, PR China
- Lorenzo Fagiano, Polytechnico di Milano, Italy
- Luis de la Torre Cubillo, UNED, Dpt. Informática y Automática, Madrid, Spain
- Luis Hernández Gómez, Technical University of Madrid, Spain
- Maarten De Vos, University of Oxford, United Kingdom
- Martin Kampel, Institute of Visual Computing & Human-Centered Technology, TUW, Austria
- Mike Dawney, Middlesex University, United Kingdom
- Moe Win, Massachusetts Institute of Technology, USA
- Mohammed Ismail Elnaggar, The Ohio State University, USA
- Nataša Vilić, Filozofski fakultet, Univerzitet u Banja Luci, BiH
- Nellie Swart, University of South Africa, Pretoria
- Nuno Gonçalo Coelho Costa Pombo, University Beira Interior, Portugal
- Nuno Manuel Garcia dos Santos, University Beira Interior, Portugal
- Özge Ercan, Faculty of Sport Sciences – Sinop University-Türkiye
- Riste Temjanovski, Goce Delčev University, Macedonia
- Roberta Grossi, Horizons University, France
- Simona Distinto, Department of Life and Environmental Sciences University of Cagliari, Italia
- Slobodan Luković, ALaRI, Switzerland
- Snezana Lawrence, Bath Spa University, United Kingdom
- Stanimir Sadinov, Technical University of Gabrovo, Bulgaria
- Vassilis S. Moustakis, Technical University of Crete, Greece
- Violeta Grubliene, Klaipeda University, Lithuania
- Vladimir Terzija, University of Manchester, United Kingdom

REVIEWERS COMMITTEE

- Mladen Veinović, Singidunum University, Serbia
- Mladan Jovanović, Singidunum University, Serbia
- Miodrag Živković, Singidunum University, Serbia
- Dragan Cvetković, Singidunum University, Serbia
- Marko Tanasković, Singidunum University, Serbia
- Zora Konjović, Singidunum University, Serbia
- Milan Tair, Singidunum University, Serbia
- Jelena Gajić, Singidunum University, Serbia
- Ivana Brdar, Singidunum University, Serbia
- Valentina Gavranović, Singidunum University, Serbia
- Srdan Marković, Singidunum University, Serbia
- Duško Lukač, Rheinische Fachhochschule Köln – University of Applied Sciences, Germany
- Gordana Pešaković, Argosy University, USA
- Nataša Vilić, Filozofski fakultet, Univerzitet u Banja Luci, BiH
- Nuno Gonçalo Coelho Costa Pombo, University Beira Interior, Portugal
- Stanimir Sadinov, Technical University of Gabrovo, Bulgaria
- Slobodan Luković, ALaRI, Switzerland
- Jovica V. Milanović, University of Manchester, United Kingdom

ORGANISING COMMITTEE

- Mladen Veinović, Singidunum University, Serbia
- Milan Tair, Singidunum University, Serbia
- Nebojša Bačanić Džakula, Singidunum University, Serbia
- Marko Tanasković, Singidunum University, Serbia
- Zora Konjović, Singidunum University, Serbia
- Jelena Gajić, Singidunum University, Serbia
- Valentina Gavranović, Singidunum University, Serbia
- Miroslav Popović, Singidunum University, Serbia
- Nina Dragičević, Singidunum University, Serbia
- Miodrag Živković, Singidunum University, Serbia
- Jelena Gavrilović, Singidunum University, Serbia
- Miloš Antonijević, Singidunum University, Serbia
- Mladan Jovanović, Singidunum University, Serbia
- Srdan Marković, Singidunum University, Serbia
- Marina Marjanović, Singidunum University, Serbia
- Lazar Dražeta, Singidunum University, Serbia
- Aleksandar Mihajlović, Singidunum University, Serbia
- Nikola Savanović, Singidunum University, Serbia
- Miloš Mravik, Singidunum University, Serbia
- Predrag Obradović, Singidunum University, Serbia
- Miloš Višnjić, Singidunum University, Serbia
- Aleksa Vidaković, Singidunum University, Serbia
- Petar Kresoja, Singidunum University, Serbia
- Teodor Petrović, Singidunum University, Serbia
- Jelena Janačković, Singidunum University, Serbia
- Bojan Papaz, Singidunum University, Serbia

INTERNATIONAL SCIENTIFIC CONFERENCE ON INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND DATA SCIENCE

Publisher: Singidunum University, 32 Danijelova Street, Belgrade
 Editor-in-Chief: Milovan Stanišić, PhD, Singidunum University, Serbia
 Prepress: Miloš Višnjić, Tijana Vučićević
 Design: Aleksandar Mihajlović
 Year: 2025
 Circulation: 200
 Printed by: Caligraph, Belgrade ISBN: 978-86-7912-841-6

Contact us:
 Singidunum University
 32 Danijelova Street, 11010 Belgrade, Serbia
 Phone No. +381 11 3093220, +381 11 3093290,
 Fax. +381 11 3093294
 E-mail: sinteza@singidunum.ac.rs
 Web: sinteza.singidunum.ac.rs

Copyright © 2025

All rights reserved. No part of this work covered by the copyright herein may be reproduced, transmitted, stored or used in any form or by any means graphic, electronic, or mechanical, including but not limited to photocopying, recording, scanning, digitizing, taping, Web distribution, information networks, or information storage and retrieval systems, without including the primary source of this Book of Proceedings.



ABOUT SINTEZA 2025

The 12th International Scientific Conference Sinteza was held on May 9, 2025, in person at the premises of Singidunum University. The conference focused on information technology, computer science, data science, and their applications in engineering systems, education, foreign language teaching, and sports. This year's program placed particular emphasis on topics related to artificial intelligence, machine learning, and data-driven research, especially in the context of solving real-world challenges.

As in previous years, the conference was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia.

Sinteza 2025 once again brought together researchers from Serbia and abroad. A total of 103 papers were submitted, out of which 78 were accepted and 25 were rejected. The conference featured 200 unique authors from 14 different countries, including Serbia. All accepted papers for the Sinteza 2025 conference are scientific papers. Additionally, all accepted papers underwent a thorough review process, including technical, linguistic, and expert evaluation, as well as iThenticate plagiarism checks and verification to assess whether, and to what extent, the text was generated by artificial intelligence tools.

At Sinteza 2025, we welcome academics, researchers, and industry professionals to share their knowledge and insights on the latest developments in Information Technology, Computer Science, and Data Science. We also recognise the importance of supporting the next generation of professionals in these fields, and we are pleased to offer a specially reserved student session at the conference.

Each parallel session was interactive and engaging, providing presenters the opportunity to share their research papers, case studies, and innovative projects. Conference participants actively engaged in discussions, offering feedback and exchanging insights with experts in the field.

We extend our sincere gratitude to the distinguished speakers of the plenary session, all participants, and members of the Scientific Committee. A special thanks goes to our colleagues from the Organising Committee for their dedicated efforts in ensuring the technical and logistical success of Sinteza 2025.

Sincerely,

Sinteza 2025 Organising Committee



CONFERENCE SPEAKERS IN PLENARY SESSION

Conference Chairmen:

Mladen Veinović, PhD – *Dean of FIR, Teaching Vice-rector, Singidunum University*

Valentina Gavranović, PhD – *Singidunum University, Serbia*

Opening Speech:

Mladen Veinović, PhD – *Dean of FIR, Teaching Vice-rector, Singidunum University*

Nebojša Bačanin Džakula, PhD – *Vice-Rector for Scientific Research, Singidunum University, Serbia*

Konjović Zora, PhD – *Singidunum University, Serbia*

Keynote speakers:

Vladan Devedžić, PhD – *Professor Academician, University of Belgrade, Faculty of Organizational Sciences Serbian Academy of Sciences and Arts, Department of Technical Sciences*

Milan Segedinac, PhD – *Professor, University of Novi Sad, Faculty of Technical Sciences*

Vladimir Milovanović, PhD – *Professor, University of Kragujevac, Faculty of Engineering, Department of Electrical Engineering and Computer Sciences, Center for Integrated Systems*



CONFERENCE KEYNOTE SPEAKERS:



Vladan Devedžić

University of Belgrade, Faculty of Organizational Sciences
Serbian Academy of Sciences and Arts, Department of Technical Sciences

Biography

Vladan Devedžić is a Professor of Computer Science and Software Engineering at the University of Belgrade, Faculty of Organizational Sciences.

He is the founder and Head of the research group focused on Artificial Intelligence (GOOD OLD AI research network). He is also the founder of the Artificial Intelligence Laboratory at his home faculty.

Since 2021, he has been a corresponding member of the Serbian Academy of Sciences and Arts (SASA) at the Department of Technical Sciences.

According to the list of world's top scientists, published by Stanford University, he is among 0.6% of the most cited researchers in the "career" category in the field of Artificial Intelligence (for the period 1996–2024).

Vladan Devedžić's continuous professional goal is to bring together ideas from the field of Artificial Intelligence /Intelligent Systems and Software Engineering. His current interests include artificial intelligence, programming education, software engineering, and technology-enhanced learning.

He has authored/co-authored about 400 research papers, published in international and national journals or presented at international and national conferences, as well as six books on intelligent systems and software engineering. Some of his papers have been selected by foreign editors and published in books on artificial intelligence systems. He has also given more than 20 invited talks, keynote talks and plenary talks at international conferences, as well as more than 20 tutorials.

His international research cooperation and collaboration record includes universities and research institutes from EU, USA, Canada, Japan, Australia, New Zealand, India and China. More recently, he has also initiated collaborations with universities in South America.

Keynote speech title

Red rAI is pouring down

Abstract

Artificial Intelligence (AI) is a permanent presence, and education is already integrating it. Teachers leverage generative AI (GenAI) to design and evaluate assignments, create interactive learning programs, produce engaging materials, and much more. Administrative staff use GenAI tools to develop and manage budgets, analyze and generate ideas, establish and modify policies, among other uses.

However, a key question remains: what AI topics should be taught, at what educational levels, and how? This question is made more difficult by the rapid developments and changes in the AI field, which creates a challenge for educators and administrators to continuously update their courses and study programs to keep pace with the field's evolution.

This presentation will illustrate this difficulty using examples from existing courses and pedagogical approaches, suggesting possible strategies for creating and maintaining AI curricula. It will also touch upon in-demand jobs in the AI industry, indicating the knowledge and skills that employers seek when advertising positions.



Milan Segedinac

University of Novi Sad, Faculty of Technical Sciences

Biography

Milan Segedinac received his M.Sc. in 2008 and Ph.D. in 2014 in Computer Science both from the Faculty of Technical sciences University of Novi Sad where he currently holds the position of a full professor in applied computer science and informatics. His research focuses on Artificial Intelligence and Software Engineering, particularly AI-driven educational technologies and knowledge representation in computer-supported education.

He has authored and co-authored over 50 research papers published in international and national journals or presented at international and national conferences, especially in the field of technology-enhanced learning. He has also participated in more than 20 commercial and scientific projects.

In addition to his academic contributions, Milan Segedinac collaborates with universities and research institutes worldwide, working on advancing AI applications in education and fostering innovation in intelligent learning systems.

Keynote speech title

Bridging Symbolic and Connectionist AI: Neuroevolutionary Approaches to Graph-Based Knowledge Representation

Abstract

Graph-based methods for knowledge representation are among the oldest techniques in AI, yet they remain crucial today, particularly in applications like knowledge graphs. At the same time, connectionist AI, based on neural networks, has seen remarkable advancements. In recent years, the boundary between symbolic and connectionist AI has become increasingly blurred, leading to the rise of neuro-symbolic AI. This convergence offers new opportunities to combine structured reasoning with learning-based approaches.

In this talk, we explore a novel method that applies neuroevolutionary techniques to the construction of knowledge spaces. Traditional approaches to building these mathematical models often struggle with scalability, making them less effective for large and complex domains. The approach that we discuss in this talk bridges the gap between connectionist and symbolic models by defining a set of analogies between knowledge spaces and neural networks, allowing evolutionary algorithms to shape an optimal structure for knowledge representation.

To illustrate the potential of this method, we focus on its application in AI-augmented learning management systems, where it helps model student learning and adapt educational pathways. By integrating neuro-evolution with graph-based knowledge representation, this approach offers a new way to enhance adaptive learning and contributes to the broader effort of bridging symbolic and connectionist AI.



Vladimir Milovanović

University of Kragujevac, Faculty of Engineering,
Department of Electrical Engineering and Computer Sciences, Center for Integrated Systems

Biography

Vladimir Milovanović received the Dipl.-Ing. degree in Electrical Engineering from the University of Belgrade, Serbia in 2005, and the Ph.D. degree from the Delft University of Technology, the Netherlands, in 2010.

Since the beginning of 2014, he was working as a Postdoctoral Scholar with the University of California, Berkeley. Before joining Berkeley Wireless Research Center, from 2011 he was with Vienna University of Technology, Austria as a Postdoctoral Research Fellow. Presently, he is holding a position of an Associate Professor with the Department of Electrical Engineering and Computer Sciences at the Faculty of Engineering, University of Kragujevac, Serbia, and serves as the managing director of the Center for Integrated Systems within the same institution.

Dr. Milovanović has also held advisory, consulting, or visiting positions with Texas Instruments, NXP Semiconductors, Infineon Technologies, Sony and Broadcom.

His research focuses and interests include design, modeling and optimization of analog, mixed-signal and digital integrated circuits and systems, along with the development and implementation of efficient artificial intelligence and signal processing algorithms.

Prof. Milovanović is the recipient of the Best Student Paper Award at the 2009 IEEE Bipolar/BiCMOS Circuits and Technology Meeting and the Best Paper Awards at the 2014 IEEE International Conference on Microelectronics and the 2024 IcETRAN.

Keynote speech title

Popular TV Quiz Shows: From Pastime to Large Language Model Benchmarking and Back

Abstract

Ever since IBM's computer system named Watson, capable of answering questions posed in natural language, outscored previous champions of the popular television game show Jeopardy! back in 2011, there has been an unprecedented rise of artificial intelligence (AI) in general. A Serbian quiz show counterpart, TV Slagalica does not just serve as viewing entertainment but is a real nursery of challenging problems and research ideas. The quiz features several games, few of which can be optimally solved even by computer science freshmen as part of their regular coursework. However, solving one of the games, named Associations, poses an ambitious task both for graduate scholars and also for the state-of-the-art large language models (LLMs).

This keynote outlines the development of an automated (intelligent) system that is able to play the Associations game on par with human players. The aspects of training data preparation, test data extraction, model assessment, and generalization will be covered. Alternative use cases, such as the ones for puzzle preparation, will be mentioned. The performance of leading-edge LLM-based chatbots that pass the Turing test, like OpenAI's ChatGPT, is thoroughly evaluated. It is demonstrated that with each successive model generation, the success rate of correct solution guesses is increased. Consequently, carefully curated test sets can be used for model benchmarking since, unlike already established benchmarks, the game of word associations is also challenging for humans.

As a conclusion, not only can it be proved that modern LLMs are close to, if not already achieving, superhuman performance in almost all TV quiz show games, but when, in the near future, foundation models start to significantly outperform humans in some areas, this benchmark can potentially help in quantifying that gap.



CONTENTS



COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE SESSION

CHAIRMAN: Miodrag Živković

3 - 9	ENHANCING RETRIEVAL-AUGMENTED GENERATION WITH GRAPH-BASED RETRIEVAL AND GENERATIVE MODELING Dejan Vujić, Angelina Njeguš, Nebojša Bačanić Džakula
10 - 15	TRAFFIC SIGN RECOGNITION USING COMPUTER VISION Vesna Radojčić, Miloš Dobrojević
16 - 22	ANT COLONY OPTIMIZATION ALGORITHM FOR FRONTEND RESOURCE PRIORITIZATION Dejan Bulaja, Katarina Stojiljković, Miodrag Živković, Nebojša Bačanić Džakula, Tamara Živković
23 - 28	CYBERSECURITY THREATS FOR MEDICAL IMAGING DEVICES A SYSTEMATIC REVIEW Ahmed Shakeel, Khalid Samina, Mehmood Yasir, Okwu O. Modestus
29 - 34	THE IDS SECURITY CHALLENGE SOLUTIONS OFFERED BY META-HEURISTIC OPTIMIZATION Dušan Cvetković, Miodrag Živković, Nebojša Bačanić Džakula
35 - 40	FEDERATED LEARNING SETTING FOR E-LEARNING COURSE RECOMMENDATIONS Miloš Jolović, Dušan Kostić, Aleksandar Joksimović, Talib Tahirović, Petar Luković
41 - 47	EFFECTS OF ADAM OPTIMIZER VARIANTS ON BRAIN TUMOR SEGMENTATION TASK Samson Offorjindu, Marina Marjanović, Timea Bezdán
48 - 54	INTELLIGENT SOFTWARE SYSTEMS FOR MULTI-TENANT CLOUD ENVIRONMENTS CHALLENGES AND SOLUTIONS Veljko Pakević, Angelina Njeguš
55 - 62	SPEED CONTROL OF A ROTARY SERVO-BASE UNIT - LYAPUNOV AND MIT RULE APPROACHES Mitra Vesović, Vladimir Zarić, Natalija Perišić, Radiša Jovanović
63 - 70	CONTROL OF A LIQUID LEVEL SYSTEM BASED ON A PROPORTIONAL-SUM CONTROLLER USING WHALE OPTIMIZER Vladimir Zarić, Natalija Perišić, Mitra Vesović, Radiša Jovanović
71 - 77	A PLATFORM-AGNOSTIC DEPLOYMENT STRATEGY LEVERAGING REAL-TIME HOST METRICS FOR OPTIMIZED RESOURCE UTILIZATION Teodor Petrović, Aleksa Vidaković, Petar Kresoja, Nikola Savanović, Saša Adamović
78 - 85	LEVERAGING LLMS FOR AUTOMATIC FORUM SCRAPER GENERATION Miloš Pavković, Jelica Protić, Petar Kresoja



INFORMATION TECHNOLOGY SESSION

CHAIRMAN: Timea Bezdán

89 - 96	CASE STUDY OF PERFORMANCE ON THE KVM HYPERVISOR-BASED VIRTUALIZATION RELATED TO NATIVE HOST Borislav Đorđević, Kristina Janjić, Nenad Kraljević
97 - 104	FILE SYSTEM PERFORMANCE COMPARISON WITH KVM AND XEN AS TYPE-1 LINUX-BASED HYPERVISORS Borislav Đorđević, Kristina Janjić, Nenad Kraljević
105 - 109	THE PROBLEMS AND SOLUTIONS DURING AGILE TEAMS MEETINGS IN IT COMPANIES – A LITERATURE REVIEW Maja Gaborov, Nada Ratković Kovačević, Dragan Kreculj, Mihajlo Mulić
110 - 116	DIGITAL WATERMARKING IN IMAGES USING DCT - EMBEDDING, EXTRACTION AND VISUAL QUALITY PRESERVATION Nenad Stanojević, Petar Spalević, Stefan Panić, Đorđe Šarčević, Srđan Mitrović
117 - 127	PERFORMANCE OPTIMIZATION OF FILE SYSTEMS FOR DOCKER CONTAINERS Sava Stanišić, Borislav Đorđević, Olga Ristić, Ivan Tot
128 - 135	ONTOLOGICAL MODELLING AND REASONING FOR THE ABSTRACTHOME CLASS OF THE BAB FRAMEWORK FOR PAIS SYSTEMS Borivoj Bogdanović, Vidan Marković, Đorđe Obradović, Milan Segedinac, Zora Konjović
136 - 143	DIGITAL INNOVATION A PILOT IN TRANSNATIONAL HYBRID TEACHING Michael Detyňa, Sue Willman, Muhammad Sami Siddiqui
144 - 149	APPLICATION PROTOTYPE FOR CENTRALIZED AND AUTOMATED NETWORK MANAGEMENT SYSTEM Petar Kresoja, Marko Šarac, Aleksa Vidaković, Teodor Petrović, Miloš Mravik
150 - 157	MULTI-CRITERIA DECISION-MAKING IN THE EVALUATION OF SOFTWARE TESTING METHODS Ana Bašić, Dejan Viduka
158 - 163	DYNAMICS OF MODERN WEB SCRAPING - CLIENT EXPECTATIONS VS TECHNICAL REALITY Umar Khalid, Armeen Shahid
164 - 171	PERFORMANCE ANALYSIS OF HAAR CASCADE-BASED FACE DETECTION IN MULTI-FACE IMAGES UNDER DIVERSE COMPRESSION ALGORITHMS Ivan Šarkoćević, Vladimir Maksimović, Branimir Jakšić, Petar Spalević, Đoko Bandur



DATA SCIENCE AND APPLICATIONS SESSION

CHAIRMAN: Tamara Živković

175 - 183	FORECASTING BASE METALS PRICES - A COMPARISON OF VARIOUS BAYESIAN-BASED METHODS Krzysztof Drachal, Joanna Jędrzejewska
184 - 192	SIMULATION OF ELECTRICAL POWER SYSTEMS OF GREECE (CRETE) USING THE INTERPSS Dionysia Svarna, Nterekas Serafeim, Georgia Ntereka Alexios
193 - 199	MANDIBLE DOSE COMPARISON BETWEEN FOTELP-VOX SIMULATIONS AND IMRT IN HEAD AND NECK RADIOTHERAPY Milena Živković, Marina Svičević, Dragana Krstić, Taha Wais Yaseen, Lazar Krstić
200 - 205	REVIEW OF META-HEURISTIC APPROACHES FOR BOOSTING MODEL OPTIMIZATION Miloš Grubješić, Miodrag Živković, Tamara Živković, Nebojša Bačanin Džakula
206 - 212	INTRUSION DETECTION TECHNIQUES AND SWARM INTELLIGENCE CYBERSECURITY REVIEW Zorana Krsmanović, Stojanka Tešanović, Aleksandar Petrović, Miodrag Živković, Tamara Živković
213 - 220	DATA SCIENCE MEETS FRAGRANCE - ANALYZING USER REVIEWS TO DECODE EMOTIONAL CONNECTIONS TO PERFUME NOTES Milena Nikolić, Marina Marjanović
221 - 227	APPLICATION OF THE CRAIRSIS AI-BASED FRAMEWORK FOR THE ANALYSIS OF PCB-170 IN HUMAN BREAST MILK Timea Bezdan, Gordana Jovanović, Andreja Stojić, Snježana Herceg Romanić, Mirjana Perišić
228 - 232	THE MANN-KENDALL TEST FOR WEATHER FORECASTING IN SMART CITIES Miloš Todorov, Ninoslava Tihi, Marko Pavlović, Suad Šuljević
233 - 240	IMPLEMENTATION OF THE DEBUGGING SUPPORT FOR THE LLVM OUTLINING OPTIMIZATION Vojislav Tomašević, Đorđe Todorović, Maja Vukasović
241 - 246	AI AND CYBERSECURITY IN E-COMMERCE - HOW AI INTEGRATION AFFECTS DATA PROTECTION Emilija Jovanović, Mladen Veinović, Jovanović Miloš



ADVANCED TECHNOLOGIES AND APPLICATIONS SESSION

CHAIRMAN: Mladen Jovanović

251 - 258	HYPERSONIC TARGET DETECTION USING EOIR-EMULATED SYNTHETIC DATA - A CASE STUDY OF X-43A Slobodan Pajić, Gerrit Viljoen, Marko Tanasković
259 - 266	UNSUPERVISED AND SEMI-SUPERVISED LEARNING TECHNIQUES IN CONTEMPORARY EDUCATIONAL APPLICATION Veljko Aleksić
267 - 273	BLOCKCHAIN-BASED PLATFORM FOR TRACKING STUDENTS' ACHIEVEMENTS Amar Mujezinović, Aleksandar Joksimović, Dušan Kostić, Miloš Jolović, Milica Simić
274 - 280	CHALLENGING DEEPSEEK-R1 WITH SERBIAN HIGH SCHOOL MATH COMPETITION PROBLEMS Nemanja Vučićević, Marina Svičević, Aleksandar Milenković
281 - 288	THE DESIGN CHARACTERISTICS OF INTELLIGENT TUTORING SYSTEMS FOR STEM EDUCATION Veljko Aleksić, Dionysios Politis
289 - 296	THE APPLICATION OF ARTIFICIAL INTELLIGENCE TO ASSIST PEOPLE WITH HEARING IMPAIRMENTS AN OVERVIEW OF EXISTING SOLUTIONS Dobroslav Drakul, Dragan Golubović, Marko Tanasković
297 - 302	DESIGN AND DEVELOPMENT OF DIGITAL TWINS IN SIMULINK Dragan Kreculj, Đorđe Dihovični, Nada Ratković Kovačević, Aleksandra Mitrović, Maja Gaborov
303 - 308	VALIDATION OF THE FDM-BASED ADDITIVE MANUFACTURING METHOD FOR RAPID PROTOTYPING USING THE EXAMPLE OF THE EIFFEL TOWER MODEL Dragan Cvetković, Nikola Vorkapić, Ivana Cvetković, Srđan Trajković, Aleksandar Cvetković
309 - 314	APPLICATION OF ADDITIVE MANUFACTURING TECHNOLOGIES IN THE PRODUCTION OF A MONASTERY MODEL FOR THE PRESERVATION OF SERBIA'S CULTURAL HERITAGE Dragan Cvetković, Nikola Vorkapić, Ivana Cvetković, Aleksandar Cvetković, Srđan Trajković



ADVANCED TECHNOLOGIES AND APPLICATIONS SESSION

CHAIRMAN: Mlađan Jovanović

315 - 322	OPEN-SOURCE AND CLOUD-BASED SOLUTIONS FOR EFFICIENT HFSWR SITE PLANNING Darko Marjanović, Dragan Golubović
323 - 330	CULTURAL TOURISM IN THE AGE OF EMERGING TECHNOLOGIES Sava Cipriana, Dragotă Tiberiu–Gabriel
331 - 336	LEARNING A FIRST PROGRAMMING LANGUAGE WITH LLM TUTOR Milić Vukojčić, Ivana Korica



MANAGEMENT AND TECHNOLOGY SESSION

CHAIRMAN: Lazar Dražeta

341 - 346	ENHANCED MONTE CARLO SCHEDULE ANALYSIS - EVALUATION OF THE OPEN-SOURCE PERT-BASED SIMULATION TOOL Goran Avlijaš
347 - 353	MICROSOFT COPILOT AS A TRANSFORMATIVE TOOL IN BUSINESS - OPPORTUNITIES AND CHALLENGES Milica Palić, Miloš Mravik, Marko Šarac
354 - 360	IDENTIFICATION OF SAFE HELICOPTER LANDING ZONES AND AN OVERVIEW OF HELIPORTS IN SERBIA Zoran Kričković, Emina Kričković
361 - 368	THE EFFECTIVENESS OF ONLINE EXECUTIVE COACHING IN LEADERSHIP DEVELOPMENT - A CASE STUDY Maja Relić, Lazar Dražeta, Borivoje Đokić
369 - 375	ETHICAL CONCERNS AND MASS AGENTIC AI ADOPTION Sonja Jovanović, Lazar Dražeta, Aleksandar Petrović, Nebojša Bačanin Džakula, Miodrag Živković
376 - 381	ARTIFICIAL INTELLIGENCE IN THE CREATIVE INDUSTRY - STRATEGIC IMPLEMENTATION OF MARKETING TOOLS Nemanja Jeličić, Ivana Vasić, Jelena Gajić, Aleksandar Mihajlović
382 - 388	THE ECONOMICS OF AI-POWERED CALL CENTER DEVELOPMENT USING CHATGPT FOR THE NEEDS OF AN AUTOMOTIVE RETAIL BUSINESS Pavle Dakić, Tjaša Heričko, Živana Kljajić, Vladimir Todorović
389 - 395	ASSESSMENT OF WASTEWATER TREATMENT, WATER QUALITY, AND CLIMATIC STRESSORS ON FRESHWATER RESOURCES IN THE EUROPEAN UNION COMPARED TO SERBIA - TRENDS AND IMPLICATIONS Aleksandra Mitrović, Nada Ratković Kovačević, Ivana Ilić, Dejan Ilić, Maja Anđelković
396 - 401	THE ROLE OF OPEN INNOVATION IN ENHANCING MANAGERIAL COMPETENCIES IN AI-DRIVEN SYSTEMS Tamara Papić, Bojan Gutić, Nina Pantelić, Nikola Petrović



INFORMATION TECHNOLOGY IN TEACHING FOREIGN LANGUAGES SESSION

CHAIRMAN: Valentina Gavranović

405 - 410	ARTIFICIAL INTELLIGENCE IN MOBILE LANGUAGE LEARNING - DUOLINGO AND THE RISE OF A NEW EDUCATIONAL ERA Tijana Gajić, Jelena Nikolić, Neda Maenza, Aleksandra Gagić
411 - 417	THE DIGITAL DOPPELGÄNGERS OF NIKOLA TESLA AND BRANISLAV NUŠIĆ - A NEW APPROACH TO INTERACTIVE LEARNING AND CULTURAL HERITAGE Milena Škobo, Milena Šović
418 - 423	AN ANALYSIS OF THE USE OF ORTHOGRAPHIC RULES IN STUDENTS' WRITTEN ASSIGNMENTS IN DIGITAL ENVIRONMENTS Saša Čorboloković, Valentina Gavranović
424 - 430	DIGITAL PLATFORMS FOSTERING INTERCULTURAL COMPETENCE IN LANGUAGE CLASSES Ivana Đerić, Neda Maenza, Nataša Stanišić, Aleksandra Gagić, Tijana Gajić
431 - 437	CHALLENGES IN USING ACTIVE LEARNING IT STRATEGIES IN FOREIGN LANGUAGE TEACHING FOR HIGHER EDUCATION Milica Pančevac, Ana Tripković
438 - 444	DISRUPTION IN COMMUNICATION IN HYBRID FOREIGN LANGUAGE TEACHING MODEL Katarina Nasradin, Miloš Pupavac, Maja Veljković Michos, Valentina Bošković Marković, Darija Lunić
445 - 450	AI-GENERATED TEXTS IN A FOREIGN LANGUAGE CLASS - PROS AND CONS Vladimir Fedorov, Miloš Pupavac, Maja Rončević



INFORMATION TECHNOLOGY IN SPORTS SESSION

CHAIRMAN: Srđan Marković

455 - 460	THE GROWING ROLE OF NUTRITION APPLICATIONS IN HEALTH AND WELLNESS - THE DEVELOPMENT OF CISHRANA, SERBIA'S FIRST NUTRITION APPLICATION Marija Anđelković, Nenad Dikić
461 - 466	PARENTS' ATTITUDES TOWARDS THE INTEGRATION OF INNOVATIVE TECHNOLOGIES IN PHYSICAL EDUCATION FOR PRESCHOOL CHILDREN Petar Nikodijević, Goran Vještica, Vladan Vođević, Aleksandar Gadžić
467 - 473	THE USAGE OF THE BLAZEPOD PLATFORM IN SPORTS Vladislav Pavlović, Vladan Marković, Miloš Milošević
474 - 480	THE PROFOUND IMPACT OF THE MOBILE APPLICATION INDUSTRY ON FITNESS Tamara Ratković, Uroš Jovanov
481 - 486	ISOMETRIC STRENGTH DIFFERENCES IN THE ATHLETIC SHOULDER TEST IN FEMALE VOLLEYBALL PLAYERS Damjana V. Čabarkapa, Vladimir Banković, Nenad Trunić, Dimitrije Čabarkapa
487 - 492	EVALUATION OF TIME MEASUREMENT METHODS – A REVIEW Srđan Marković, Slađanja Rakić
493 - 498	THE PRESENCE OF TECH COMPANIES IN SUPER BOWL ADS FROM 2021 TO 2025 Aleksandra Belačić, Slavko Alčaković



STUDENT SESSION

CHAIRMEN: Milan Paroški, Aleksandar Petrović

503 - 510	UTILIZING BUSINESS ANALYTICS, BIG DATA, AND VISUALIZATION FOR SALES PERFORMANCE OPTIMIZATION Aleksandar Radivojević, Miloš Mravik, Marko Šarac
511 - 517	WEB APPLICATION FOR SCHEDULING DOCTOR APPOINTMENTS Bojan Papaz, Miloš Mravik, Marko Šarac
518 - 523	THE IMPACT OF ARTIFICIAL INTELLIGENCE IN MARKETING Anđela Pavlović, Marko Šarac
524 - 530	A COMPARATIVE STUDY OF OBJECT DETECTION ALGORITHMS FOR SECURITY APPLICATIONS Roman Kriuchkov, Timea Bezdán
531 - 535	HANDWRITTEN DIGIT RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS AND BIG DATA PROCESSING Pavle Matijašević, Miloš Mravik
544 - 549	SDLC-INDEPENDENT PYTHON-BASED QUERY PERFORMANCE BENCHMARKING APPROACH AND PRACTICAL OPTIMAL DATABASE SELECTION GUIDELINES Katarina Milojković, Petar Spalević, Nikola Vasić, Nikola Milojković, Hristina Milojković
419 - 424	MULTIMODAL RETRIEVAL-AUGMENTED GENERATION IN KNOWLEDGE SYSTEMS - A FRAMEWORK FOR ENHANCED SEMANTIC SEARCH AND RESPONSE ACCURACY Marko Mihajlović
550 - 554	OPTIMIZATION OF KUBERNETES - RESOURCE ALLOCATION AND DYNAMIC SCALING Džemil Sejdija, Aldina Avdić
555 - 560	GAME DEVELOPMENT – HOW TO CREATE AN IMMERSIVE STORYTELLING EXPERIENCE Nikola Peruničić, Mlađan Jovanović
561 - 566	THE ROLE OF ARTIFICIAL INTELLIGENCE IN WASTEWATER TREATMENT Milica Radojković, Milena Rikalović



COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE SESSION





ENHANCING RETRIEVAL - AUGMENTED GENERATION WITH GRAPH-BASED RETRIEVAL AND GENERATIVE MODELING

Dejan Vujić*,
[0009-0008-2037-5068]

Angelina Njeguš,
[0000-0001-8682-7014]

Nebojša Bačanić Džakula
[0000-0002-2062-924X]

Singidunum University,
Belgrade, Serbia

Abstract:

This paper presents the design and implementation of a robust Retrieval-Augmented Generation (RAG) system that integrates advanced retrieval, ranking, and generative techniques to address knowledge-intensive tasks. The system combines dense retrieval using ChromaDB, metadata-driven keyword extraction with YAKE and KMedoids algorithm for clustering keywords, graph-based retrieval leveraging PageRank, and cross-encoder re-ranking to deliver precise and contextually relevant results. These retrieval outputs are synthesized into high-quality conversational responses using Hugging Face models and Google API. A modular pipeline ensures scalability, seamlessly integrating various retrieval and generative components. Evaluation results demonstrate high retrieval precision, improved recall through graph-based methods, and enhanced response quality through structured prompt engineering. This work highlights the effectiveness of combining diverse techniques in RAG systems, offering a foundation for scalable, reliable, and context-aware applications in domains such as customer support, education, and research.

Keywords:

Retrieval-Augmented Generation, Dense Retrieval, Re-Ranking, Graph-Based Retrieval Keywords, Generate Modeling.

INTRODUCTION

The rapid growth of artificial intelligence (AI) has brought about significant advancements in natural language processing (NLP). Among the most transformative technologies is Retrieval-Augmented Generation (RAG), a hybrid approach that combines the strengths of retrieval systems with generative models. By integrating these paradigms, RAG enables more accurate and contextually relevant responses, addressing limitations inherent in traditional generation-only or retrieval-only methods. As highlighted in recent studies, RAG has proven effective in improving accuracy and contextuality in tasks such as question-answering and conversational agents [1].

RAG finds applications across diverse domains, including conversational agents, knowledge base construction, and personalized content generation. For instance, OpenAI's and Meta's research shows that augmenting generative language models with external retrieval improves factual consistency and reduces hallucinations in generated outputs [2].

Correspondence:

Dejan Vujić

e-mail:

dejan.vujic.24@singimail.rs





The intersection of retrieval and generation poses unique challenges and opportunities for software engineering. Implementing a robust RAG system requires careful consideration of data management, model integration, and performance optimization. For example, designing pipelines to handle large-scale retrieval while maintaining low latency is critical [3]. Similarly, advancements in transformer architectures and attention mechanisms have been instrumental in optimizing RAG-based workflows [4].

This paper discusses a Python-based RAG project's architecture, underlying models, and implementation details. By providing a detailed exploration of the project, this work contributes to understanding how retrieval and generative systems can be effectively combined, offering valuable insights for AI researchers and software engineers.

2. RETRIEVAL-AUGMENTED GENERATION OVERVIEW

Retrieval-Augmented Generation (RAG) is a hybrid framework that integrates information retrieval with text generation to tackle knowledge-intensive tasks. The approach was first introduced by [1], who demonstrated its ability to improve the quality of generated text by incorporating relevant retrieved documents into the generation process. Unlike traditional generative models that rely solely on pre-trained parameters, RAG accesses external knowledge sources, such as document collections or databases, enabling more factual and contextually accurate outputs.

This framework leverages two key components: a retriever and a generator. The retriever identifies the most relevant documents from a knowledge base, while the generator incorporates the retrieved documents to produce responses. Recent advancements in dense retrieval methods, such as Dense Passage Retrieval (DPR) [3], have further enhanced the retrieval component by enabling semantic matching of queries and documents in vector space.

RAG has been compared to other architectures that address knowledge-intensive tasks. These include:

- Open-Domain Question Answering (ODQA): Systems like DrQA [5] rely on retrieval followed by extractive reading, limiting their ability to generate free-form responses. In contrast, RAG's generative component allows for more nuanced and diverse outputs.

- Knowledge Graph-Based Systems: Approaches leveraging structured knowledge graphs (KGs) provide accurate responses by querying graph nodes. While effective for tasks requiring structured data, KGs often lack the coverage and scalability of unstructured text retrieval used in RAG [6].
- Memory-Augmented Neural Networks: Models like Memory Networks [7] and Neural Turing Machines [8] integrate memory for knowledge storage. These approaches, while powerful, are constrained by the size and scope of the memory, making RAG's ability to query external sources more scalable.

RAG's ability to dynamically query and generate content has significantly advanced state-of-the-art performance in tasks such as open-domain question answering [1], dialogue systems [2], and document summarization [4]. The combination of dense retrieval and transformer-based generation has set a new benchmark, bridging the gap between static knowledge representation and dynamic generation.

3. RESEARCH BACKGROUND

In recent years, Retrieval-Augmented Generation (RAG) has emerged as a powerful framework that enhances the capabilities of generative models by integrating external knowledge sources into the generation process. This approach addresses the limitations of purely generative models, such as hallucinations and factual inaccuracies, by retrieving relevant context from large-scale knowledge bases. Among various advancements in this domain, graph-based retrieval techniques have gained significant attention due to their ability to capture complex relationships within data, providing richer and more relevant context for generation tasks.

GRAG (Graph Retrieval-Augmented Generation) was developed to enhance both the retrieval and generation processes by emphasizing subgraph structures and maintaining awareness of graph topology to generate contextually coherent responses. The framework has demonstrated superior performance over existing RAG methods in multi-hop reasoning tasks on textual graphs, effectively mitigating hallucinations and improving response quality. [9]

The potential of Large Language Models (LLMs) for materials design has been demonstrated through the integration of retrieval-augmented ontologic graphs and



multi-agent strategies. This approach supports engineering analysis and knowledge generation by leveraging structured retrieval mechanisms, facilitating effective information retrieval and code generation for simulation purposes. [10]

A comprehensive overview of Graph RAG methodologies has been presented, formalizing workflows and discussing core technologies, applications, and future research directions. These advancements highlight the potential of Graph RAG to enhance LLM outputs by leveraging structural information in graphs, leading to more accurate and context-aware responses. [11]

A graph-driven generative model has been proposed to integrate semantic and neighborhood information for optimizing document retrieval. This approach effectively addresses the need for fast retrieval and a small memory footprint. Experimental results demonstrate superior performance over state-of-the-art methods in document hashing, preserving both semantic and neighborhood information in retrieval tasks. [12]

Large Generative Graph Models (LGGMs) were introduced, trained on a large corpus of graphs from diverse domains, enabling zero-shot generative capabilities and text-to-graph generation. These models outperformed existing methods in generating graphs across various domains and successfully integrated language model knowledge for fine-grained control over generated content. [13]

A permutation-invariant approach to graph modeling using score-based generative modeling has been developed. This approach addresses challenges in learning generative models for graph-structured data and achieves better or comparable results to existing models on benchmark datasets. These findings underscore the effectiveness of permutation-invariant methods in enhancing the accuracy and scalability of graph generation models. [14]

4. RESEARCH METHODOLOGY

4.1. PROJECT OVERVIEW

This project presents a Retrieval-Augmented Generation (RAG) system designed to process user queries by combining multiple advanced methods, such as dense retrieval, keyword-based metadata enrichment, graph-based retrieval, and re-ranking mechanisms. The system ensures high-quality, contextually enriched responses by integrating these retrieval strategies with generative modeling using the Google API.

The system addresses knowledge-intensive tasks by dynamically retrieving and processing relevant information from a knowledge base, combining structured metadata (keywords, document scores) with semantic embeddings for enhanced accuracy. Re-ranking and graph-based retrieval complement the dense retrieval pipeline by refining results and surfacing relevant content that may be indirectly linked to the user query. These retrieval outputs are subsequently utilized to construct structured prompts for generative modeling.

The system operates in the following stages:

- **Input Processing and Preprocessing:** Queries and input text are pre-processed into manageable chunks using a sentence splitter. The extracted text is tokenized to meet the constraints of the models and enriched with metadata, such as keywords derived from YAKE and KMedoids algorithm for clustering keywords. This metadata ensures semantic enrichment for downstream retrieval and ranking tasks.
- **Dense Retrieval Using ChromaDB:** Dense embeddings, generated from pre-trained transformer models, are indexed into ChromaDB, a high-performance vector database. These embeddings enable semantic similarity matching, allowing the system to retrieve top candidate documents.
- **Re-Ranking with a Cross-Encoder:** A cross-encoder model refines the initial retrieval results by directly evaluating the semantic alignment between the user query and candidate documents. This re-ranking step prioritizes the most relevant content.
- **Graph-Based Retrieval Using PageRank:** A graph-based retrieval component complements dense retrieval by leveraging relationships between documents and keywords. This graph, constructed using NetworkX, applies PageRank to identify query-specific relevance scores, uncovering additional documents with indirect relationships to the query.
- **Generative Modeling:** Retrieved documents and metadata are integrated into a structured prompt, guiding the Google API to produce conversational and contextually accurate responses. The generative model enriches user interaction by synthesizing retrieved content into coherent outputs.

The RAG system's modularity ensures scalability and adaptability across domains, while its use of multiple retrieval methods combined with advanced generative modeling enhances the precision, recall, and quality of responses.



4.2. ARCHITECTURE AND PROJECT FLOW

The Retrieval-Augmented Generation (RAG) system integrates several advanced techniques, including keyword extraction, embedding-based retrieval, re-ranking with a cross-encoder, graph-based retrieval using PageRank, and generative modeling. This section outlines the technical details and tools used to implement these components.

Preprocessing and Metadata Enrichment

The preprocessing phase begins by splitting long text into smaller, semantically coherent chunks using a sentence splitter and tokenizer. Extracted text chunks are enriched with metadata, including keywords identified using the YAKE library and KMedoids for clustering keywords. These keywords provide a semantic summary of the content, aiding both retrieval and re-ranking processes.

Semantic Retrieval

The retrieval component utilizes ChromaDB for storing and querying dense embeddings of text chunks. Dense embeddings are generated using the jinaai/jina-embeddings-v3 pre-trained model from Hugging Face, capturing the semantic structure of the text. ChromaDB indexes these embeddings, enabling rapid similarity searches based on user queries. When a query is issued, its embedding is computed and compared to the indexed embeddings in ChromaDB. The top matches are retrieved as candidates for further processing.

Re-Ranking with Cross-Encoder

The initial retrieval results are refined using a cross-encoder re-ranking model. The project employs a cross-encoder named corrius/cross-encoder-mmarmcoMiniLMv2-L12-H384-v1, which evaluates query-document pairs to compute relevance scores. These scores allow the system to prioritize documents that are most semantically aligned with the query. Re-ranking is conducted by:

- Pairing the query with each retrieved document.
- Using the cross-encoder to predict relevance scores for all pairs.
- Sorting the documents based on these scores to identify the top-ranked items.

Graph-Based Retrieval

A graph-based retrieval mechanism complements the re-ranking process by leveraging a graph representation of the knowledge base. The graph is constructed using NetworkX, where (a) nodes represent documents or keywords, and (b) edges indicate semantic relationships between nodes, derived from co-occurrence or embedding similarity. Using the PageRank algorithm, the system computes a score for each node based on its connections and relevance to the query. The personalized PageRank implementation considers query-specific weights to adjust the importance of nodes dynamically. If relevant results are found in the graph-based retrieval step, they are combined with the re-ranked documents to form a unified set of candidates.

Combining Re-Ranking and Graph Retrieval Results

The outputs of re-ranking and graph-based retrieval are merged to create a final ranked list of documents. This involves combining scores from both methods, normalizing them for consistency, and ensuring diversity in the selected documents. The combined results serve as input for the generative modeling stage.

Generative Modeling

The generative component uses Google's gemini-2.0-flash-exp Large Language Model to produce conversational responses. The process involves:

- A structured prompt is created using the retrieved and re-ranked documents, along with their metadata (e.g. keywords, URLs).
- The prompt includes detailed instructions to guide the generative model in producing accurate and user-friendly responses.
- The Gemini model generates outputs that are conversational, contextually aware, and enriched with information from the retrieved documents.
- In cases where the retrieval process yields low-confidence results, the model is instructed to inform the user of insufficient information rather than producing speculative responses.

Validation and Scoring

The system employs multiple layers of validation to ensure response quality:

- Retrieval scores from ChromaDB.
- Relevance scores from the cross-encoder.
- Graph-based PageRank scores.



Documents that meet or exceed a predefined confidence threshold are included in the prompt, ensuring factual accuracy and contextual relevance.

5. EVALUATION AND RESULTS

To validate the performance of the RAG system, a comprehensive evaluation was conducted across several dimensions, including retrieval accuracy, re-ranking effectiveness, graph-based retrieval contributions, and the quality of generative outputs. The evaluation involved testing the system on a set of predefined queries and corresponding ground-truth answers.

5.1. EXPERIMENTAL SETUP

The system was tested using a dataset comprising domain-specific text documents, metadata, and queries. The documents were preprocessed into chunks and enriched with keywords extracted using the YAKE library and KMedoids for clustering keywords. Embeddings were generated for both text chunks and keywords using pre-trained transformer models.

For retrieval and re-ranking:

- ChromaDB was employed to index document embeddings and retrieve the top 20 candidates for each query based on semantic similarity.
- The cross-encoder model was used to re-rank the initial candidates, selecting the top 10 most relevant documents.
- Graph-based retrieval with PageRank identified additional candidates, using the NetworkX library to compute relevance scores based on query-specific personalization.

Generative modeling was performed using the Google API, which produced conversational responses for each query. The responses were evaluated for accuracy, coherence, and contextual relevance.

The evaluation employed the following metrics:

- Retrieval Precision: The proportion of retrieved documents that were relevant to the query.
- Re-Ranking Effectiveness: Measured by comparing the rank order of documents before and after re-ranking using normalized Discounted Cumulative Gain (nDCG).
- Graph-Based Retrieval Contribution: Assessed by the incremental improvement in retrieval precision and recall when incorporating graph-based results.

- Generative Output Quality: Evaluated using human feedback and automated metrics, including BLEU (for linguistic similarity) and ROUGE (for content overlap).

5.2. RESULTS AND ANALYSIS

The project results show that RAG, using generative modeling and assessing the accuracy of the selected chunk based on the question asked, keywords, and the relevant retrieved chunk obtained, produces satisfactory results with the help of LLM. In Listing 1 shows an example of using RAG:

- Retrieval Performance:
 - ChromaDB achieved a high initial precision of 82%, retrieving relevant documents within the top 20 candidates.
 - The inclusion of graph-based retrieval boosted recall by 8%, highlighting its utility in capturing semantically related but less directly matched documents.
- Re-Ranking Effectiveness:
 - The cross-encoder model significantly improved the relevance of top-ranked documents, with an nDCG increase of 12% over the initial retrieval results.
 - The top 10 re-ranked documents consistently aligned with the ground-truth relevance scores.
- Graph-Based Retrieval:
 - PageRank effectively identified additional relevant documents, particularly for queries involving keywords with multiple associations in the knowledge graph.
 - Queries with high keyword density benefited the most, showing a 15% improvement in retrieval precision when combining re-ranking and graph-based results.
- Generative Output Quality:
 - BLEU and ROUGE scores indicated strong alignment between the generated responses and ground-truth answers, with an average BLEU score of 0.74 and a ROUGE-L score of 0.68.
 - The fallback mechanism ensured that low-confidence responses were flagged appropriately, maintaining user trust.



<QUERY>

What is the eco rating in A1?

<RETRIEVED TOP CHUNK>

Belgrade, July 11, 2023 - A unique methodology for assessing the impact of mobile phones on the environment is coming to Serbia. A1 Serbia is the first company in Serbia to introduce Eco Rating - an international methodology for assessing the level of sustainability and environmental impact of mobile phones from its range. As a "Network that loves the planet", A1 thus joins leading European telecommunications companies in the mission to help users make more sustainable decisions when purchasing mobile devices and encourage manufacturers to reduce the negative environmental impact of devices. To further support users to make a positive change, A1 announces an extended warranty of selected phones with a higher Eco Rating for 3 years, starting in August. The Eco Rating methodology evaluates the entire life cycle of mobile phones - from the raw materials used to make the device to the possibility of its disposal after use and gives a precise assessment of its impact on the environment. The final rating includes more than 10 analyzed categories, including energy efficiency and the amount of natural resources used in its production, the length of the device's working life, and the possibility of repair and recycling. The maximum score a phone can get is 100, and the higher the score, the better the device's performance in terms of environmental sustainability. The Eco Rating of mobile phones from the A1 range will be available in the coming weeks on the official website of this company. "Since arriving on the Serbian market, A1 has been a pioneer of innovation and positive changes in the telecommunications industry. At a time when caring for the environment is a common mission, it is our obligation to provide users with access to all information, including information about the environmental impact of the mobile phones they use and let's additionally support their ecological choices. Eco Rating brings transparency and a unique opportunity for users to recognize and compare the ecological sustainability of mobile phones and thereby personally contribute to the preservation of the environment" - said Dejan Turk, general director of A1 Serbia and A1 Slovenia.

<METADATA>

KEYWORDS: ['Eco Rating', 'environment', 'Rating', 'Eco', 'phone', 'device', 'Rating mobile', 'impact of mobile']

URL: [https://a1.rs/press-centar/a1-srbija-predstavlja-eco-rating-telefona]

<RE-RANK SCORE>

8.517

<GENERATIVE ANSWER>

Eco Rating is a unique methodology for assessing the impact of mobile phones on the environment, which A1 Serbia was the first to introduce in Serbia. This methodology evaluates the entire life cycle of the phone, from raw materials to disposal, and assesses its impact on the environment. The goal is to help users make more sustainable decisions when buying phones and encourage manufacturers to reduce the negative impact of devices.

You can find more information at <https://a1.rs/press-centar/a1-srbija-predstavlja-eco-rating-telefona>.

Listing 1. Example of Retrieval & Generate Modeling (Translated to English)

5.3. LIMITATIONS AND IMPLICATIONS

While the system performed well overall, a few limitations were observed:

- The reliance on embedding similarity in ChromaDB may miss highly relevant documents with lower semantic similarity scores.
- The graph-based retrieval approach is computationally intensive for large graphs, requiring optimization for scalability.
- Generative responses are occasionally overgeneralized, particularly for queries lacking sufficient contextual information in the retrieved documents.

The results demonstrate the effectiveness of integrating ChromaDB retrieval, cross-encoder re-ranking, and graph-based PageRank for improving retrieval accuracy and response quality. The use of Google API further enhances the system's ability to generate high-quality conversational responses.



6. CONCLUSION

This project demonstrates the development and implementation of a sophisticated Retrieval-Augmented Generation (RAG) system, integrating cutting-edge retrieval methods, ranking algorithms, and generative AI to address knowledge-intensive tasks effectively. By combining dense retrieval with graph-based approaches and re-ranking mechanisms, the system achieves both high precision and enhanced recall. Its modular design and use of advanced generative modeling through the Google API underscore its adaptability and scalability.

The system's pipeline showcases how diverse techniques can be unified to create a robust solution. The integration of ChromaDB for dense vector-based retrieval ensures fast and scalable document access, while graph-based retrieval using PageRank captures indirect relationships between documents and queries. Re-ranking with a cross-encoder further refines the results, prioritizing relevance, and quality. These retrieval outputs, combined with structured prompts, enable the generative model to synthesize contextually rich and accurate responses.

The success of this system highlights the effectiveness of combining multiple retrieval methods and generative AI in building knowledge-intensive applications. Its modular architecture makes it adaptable for various domains, such as customer support, education, healthcare, and research, where reliable and context-aware responses are crucial.

REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2020.
- [2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riede and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive Dialogues," in *Proceedings of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand, 2021.
- [3] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020.
- [4] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2021.
- [5] D. Chen, A. Fisch, J. Weston and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," in *Association for Computational Linguistics (ACL)*, Vancouver, Canada, 2017.
- [6] P. Fabio, R. Tim, L. Patrick, B. Anton, W. Yuxiang, M. H. Alexander and R. Sebastian, *Language Models as Knowledge Bases?*, 2019.
- [7] J. Weston, S. Chopra and A. Bordes, "Memory Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New York, USA, 2015.
- [8] A. Graves, G. Wayne and I. Danihelka, "Neural Turing Machines," in *Google DeepMind*, London, UK, 2014.
- [9] Y. Hu, Z. Lei, Z. Zhang, B. Pan, C. Ling and L. Zhao, "GRAG: Graph Retrieval-Augmented Generation," in *Department of Computer Science, Emory University*, Atlanta, GA 30322, USA, 2024.
- [10] M. J. Buehler, "Generative retrieval-augmented ontologic graph and multi-agent strategies for interpretive large language model-based materials design," in *ACS Engineering Au*, 2023.
- [11] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang and S. Tang, "Graph Retrieval-Augmented Generation: A Survey," 2024.
- [12] O. Zijing, S. Qinliang, Y. Jianxing, L. Bang, W. Jingwen, Z. Ruihui, C. Changyou and Z. Yefeng, "Integrating Semantics and Neighborhood Information with Graph-Driven Generative Models for Document Retrieval," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, p. 2238–2249.
- [13] Y. Wang, R. A. Rossi, N. Park, H. Chen, N. K. Ahmed, P. Trivedi, F. Dernoncourt, D. Koutra and T. Derr, "Large Generative Graph Models," 2024.
- [14] C. Niu, Y. Song, J. Song, S. Zhao, A. Grover and S. Ermon, "Permutation Invariant Graph Generation via Score-Based Generative Modeling," *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020*, Palermo, Italy, vol. 108, 2020.



TRAFFIC SIGN RECOGNITION USING COMPUTER VISION

Vesna Radojčić*,
[0000-0001-7826-1081]

Miloš Dobrojević
[0000-0003-3798-312X]

Sinergija University,
Bijeljina, Bosnia and Herzegovina

Abstract:

This paper presents a comprehensive study on the development of a traffic sign recognition system based on computer vision, which is crucial for improving road safety and supporting intelligent transport systems (ITS). By using advanced image processing techniques and machine learning algorithms, this research aims to create a reliable system capable of accurately recognizing and classifying various traffic signs. The methodology includes collecting a diverse set of traffic sign images, applying image enhancement techniques, and utilizing deep learning models for precise recognition in different environmental conditions. Experimental results indicate the system's high accuracy in recognizing traffic signs, even in cases with complex backgrounds, demonstrating its potential for integration into autonomous vehicles and advanced driver assistance systems (ADAS). This research contributes to efforts aimed at enhancing road safety and improving traffic management through automated traffic sign recognition.

Keywords:

Traffic Signs, Computer Vision, Machine Learning, Traffic Safety.

INTRODUCTION

Traffic sign recognition (TSR) is an important part of advanced driver assistance systems and autonomous vehicles, helping to improve road safety. [1]. It enables vehicles to recognize and understand signs like speed limits, warnings about children nearby, and upcoming turns [2]. Accurate recognition of traffic signs is especially significant in the context of the development of autonomous vehicles, which must be able to recognize and appropriately respond to various traffic situations in real time. The goal of this paper is to develop a method that can precisely recognize different traffic signs, such as mandatory direction signs and no parking signs. Traffic signs are recognized by their specific colors and shapes, which distinguish them from other objects in the environment, and these properties facilitate their identification in complex conditions.

Correspondence:

Vesna Radojčić

e-mail:

vradojcic@sinergija.edu.ba





Integration with machine learning and deep learning has enhanced the accuracy and efficiency of traffic surveillance, enabling traffic flow monitoring, anomaly detection, and license plate recognition [3]. Deep neural networks have shown exceptional results in recent years in various applications, including object detection and automatic traffic sign recognition [4]. However, challenges such as changing weather conditions, varying angles of observation, changing light conditions, and the variety of sign types make precise identification very challenging, especially in real time.

2. RELATED WORK

This section reviews previous research on traffic sign detection conducted in different parts of the world. Luo et al. introduced a three-stage, information-driven framework for detecting image-oriented and text-oriented signs using a vehicle-mounted camera. Despite its innovative approach, the framework's extensive post-processing stage presented a notable limitation [5]. The study "Traffic Sign Recognition Using CNN" explores the application of convolutional neural networks (CNNs) for accurately recognizing traffic signs. The proposed model leverages deep learning techniques to identify and classify traffic signs efficiently, addressing challenges like variations in size, shape, and environmental conditions. The results demonstrate the potential of CNNs in improving traffic sign recognition systems [6]. U. Venkateshwarulu and Prof. B. Manjunath, in their study "*CNN Model for Traffic Sign Recognition*," propose a convolutional neural network (CNN)-based approach for traffic sign detection and classification. Their model addresses challenges such as varying lighting conditions and occlusions, demonstrating improved accuracy and reliability in traffic sign recognition systems [7]. The study "Small Traffic Sign Recognition Method Based on Improved YOLOv7" introduces an enhanced YOLOv7 model specifically designed for recognizing small traffic signs. By optimizing the network structure and incorporating advanced detection techniques, the authors address challenges related to the detection of small and distant signs, achieving improved recognition accuracy and performance [8]. Rahul Patil [9], in his paper "Real-Time Traffic Sign Detection and Recognition System Using Computer Vision and Machine Learning," introduces a system that integrates computer vision and machine learning for real time traffic sign detection and recognition. The approach focuses on achieving high speed and accuracy, making it suitable for applications in advanced driver assistance systems and autonomous vehicles.

3. METHODS AND MATERIALS

The problem of traffic sign recognition encompasses two key components: detection and classification. Detection refers to locating the traffic sign within an image or video, while classification involves determining the type or category of the sign. Both processes are essential for the successful application of computer vision in recognizing traffic signage [10]. Signs are typically recognized through digital images captured by cameras on vehicles, traffic cameras, or other devices. Recognition can be quite complex because signs vary in position, size, and angle, and are often partially obscured by other objects or illuminated under different weather conditions [11]. To address this challenge, various techniques are used, including shape and color-based classification, feature detection, deep learning, and image segmentation methods. The goal is to achieve high accuracy in recognizing traffic signs, thereby enhancing road safety and reducing traffic accidents. These techniques are often combined to improve precision, with the choice of appropriate technology depending on the specific conditions of traffic sign recognition, available data, and computational resources.

3.1. DATASET

In preparing for this study, data collection involved a diverse set of images captured from real-life traffic situations, ensuring a comprehensive representation of various traffic scenarios. A total of 200 images were included in the dataset. The images were selected to cover a range of conditions, such as different angles, lighting situations, and sign types, allowing for a robust evaluation of the system's performance. This dataset was used to assess the effectiveness of the traffic sign detection system and its applicability in real-world environments.

3.2. IMAGE PROCESSING

Digital image processing has evolved significantly since its inception in the 1960s at Bell Laboratories [12]. It is a crucial component in the development of systems like traffic sign recognition, where the goal is to extract meaningful information from digital images or video frames. Image preprocessing techniques significantly contribute to enhancing the accuracy and efficiency of deep learning models for image classification tasks [13]. It involves various techniques and algorithms designed to enhance image quality, detect features, and classify objects accurately. This step focuses on preparing raw images for further analysis.



3.2.1. Loading an image

When loading an image, the first step is to import the required library. Then, the `cv2.imread()` function is used to load the image, storing the result in the variable `img`. This function accepts two parameters: the image file path and an integer specifying whether the image should be loaded in color or grayscale (stored in `img_1`). To display the image in a new window, the `cv2.imshow()` function is called, which requires the window name and the image variable as parameters [14]. The `waitKey()` function is then executed to pause the program until a key is pressed. Once a key is detected, the `destroyAllWindows()` function is invoked to close all open windows.

3.2.2. Converting an RGB image to grayscale

Grayscale in image processing refers to representing an image using varying shades of gray, ranging from black to white, with adjustable intensities. To create a grayscale image, you can either load the file directly in grayscale mode or, if an RGB image is already loaded,

convert it to grayscale using the `cvtColor` method from the OpenCV library [15]. In a grayscale image, each pixel is assigned a value indicating its brightness, where 0 represents black and 255 represents white in an 8-bit image. Converting an image to grayscale can also be done using the PIL (Python Imaging Library) or Pillow libraries [16].

3.2.3. Edge detection

Edge detection is a fundamental technique in image processing and computer vision, with various methods available for extracting object contours and boundaries [17]. The contours of an object can be used to define edges, which serve as fundamental elements of an image. Various algorithms are available for detecting edges, and this paper provides a comprehensive analysis of several edge detection methods, including Prewitt, Sobel, Canny, Roberts, Laplacian of Gaussian, and others.



Figure 1. Sample images from the dataset



Figure 2. Original loaded image



Figure 3. Grayscale Image Format



Figure 4. Edge Detection in Image Processing



Figure 5. Edge Detection in Image Processing



3.2.4. Noise removal

Image restoration focuses on enhancing the quality of an image by eliminating or minimizing various types of degradation, such as blur, noise, or compression artifacts. Python provides multiple libraries for image processing, including OpenCV, scikit-image, and Pillow, which can be leveraged for image restoration tasks. One of the most common restoration techniques is image denoising, which aims to reduce noise in an image, thereby improving its clarity and overall visual quality.

In conclusion, Python provides a robust set of libraries and tools for image processing, establishing it as a versatile and powerful language for a wide range of image-related tasks. By combining these techniques, image processing enables efficient and accurate recognition of traffic signs, which plays a pivotal role in improving road safety and supporting automated traffic management systems.

4. RESULTS AND DISCUSSION

First, the image is loaded. The system then applies image enhancement techniques and uses a pre-trained deep learning model to recognize traffic signs. After detecting the signs, bounding boxes are drawn around them to highlight the findings. The processed image with traffic signs is displayed to the user. The following is the pseudocode of the application:

Table 1 presents the results of traffic sign detection. The table lists different traffic sign types, along with the total number of signs and the number of signs that were successfully recognized by the system.

Based on the results, the system successfully recognized 168 out of 200 traffic signs, achieving an accuracy of 84%. These results indicate a high level of system efficiency in real-world conditions, despite variations in lighting, angles, and surroundings.

```
Load the image
  If the image is not successfully loaded:
    Display the message "Unable to load the image"
    Exit the program
Preprocess the image using image enhancement techniques
Load the pre-trained deep learning model for traffic sign detection
Detect traffic signs in the image using the model
For each detected traffic sign:
  a. Draw a rectangle around the traffic sign
Display the image with the detected traffic signs
Wait for the user to press any key
Close the image window
```

Listing 1. Pseudo code of the application

Table 1. Traffic Sign Detection Results

Traffic Sign Type	Total Signs	Recognized Signs
Speed Limit 30 km/h	25	22
Speed Limit 50 km/h	20	15
Mandatory Right Turn	18	15
Mandatory Left Turn	18	14
Mandatory Go Straight	15	13
No Left Turn	15	12
No U-Turn	14	12
Stop Sign	24	21
Marked Pedestrian Crossing	15	13
One-way Traffic Prohibition	19	16
Two-way Traffic Prohibition	17	15
Total	200	168



5. CONCLUSION

The traffic sign recognition system developed in this study represents a significant step toward safer and more efficient traffic management, providing drivers with timely alerts on key traffic signals, thereby reducing the risk of errors and enhancing overall safety. Such systems can be applied to older car models that lack advanced built-in technologies, offering a practical and affordable option for improving road safety. This is particularly relevant in Bosnia and Herzegovina, where the average age of vehicles exceeds 20 years. To make the system even more reliable and ready for broader application, further improvements are necessary, especially in adapting to different weather conditions, lighting, and complex traffic scenarios. Future research should focus on optimizing recognition models and refining algorithms to fit specific traffic contexts. In this way, traffic sign recognition technology can reach its full potential, becoming a key component of advanced driver assistance systems and modern transportation systems, providing significant benefits to all traffic participants while enhancing road safety.

REFERENCES

- [1] K. Vyshanavi and K. Guddad, "Machine Vision Based Traffic Sign Detection Methods: Review," in *Engineering, Computer Science*, 2020.
- [2] T. Kumaravel, E. Veerappampalayam, N. Sathishkumar, Premalatha, S. Dharanesh, G. Hariharan and Krishnamoorthy, "Indian traffic sign detection and recognition using deep learning," *Applied and Computational Engineering*, Vols. DOI: 10.54254/2755-2721/30/20230062, pp. 12-17, 2024.
- [3] V. Radojcic, A. S. Cvetkovic, M. Dobrojevic, P. Spalevic, and M. E. G. Jalal, "Advancements in Computer Vision Applications for Traffic Surveillance Systems," in *Sinergija University Scientific Conference with International Participation*, <https://doi.org/10.7251/zrsng2324011r>, 2023.
- [4] A. K. Muneeb, P. Heemin, and C. Jinseok, "A Lightweight Convolutional Neural Network (CNN) Architecture for Traffic Sign Recognition in Urban Road Networks," *Electronics*, p. 10.3390/electronics12081802, 2023.
- [5] L. H., Y. Y., T. B., F. Wu and F. B., "Traffic Sign Recognition Using a Multi-Task Convolutional Neural Network," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1100-1111, 2018.
- [6] J. A. Ahmed, "Traffic Sign Recognition Using CNN," *Journal of Advances in Artificial Intelligence*, pp. 19-39 DOI: 10.18178/JAAI.2025.3.1.19-39, 2025.
- [7] V. U. and M. Prof. B., "CNN MODEL FOR TRAFFIC SIGN RECOGNITION," *EPRA International Journal of Research & Development (IJRD)*, pp. 130-136 DOI:10.36713/epra18355, 2024.
- [8] B. Meng and W. Shi, "Small Traffic Sign Recognition Method Based on Improved YOLOv7," pp. DOI: 10.21203/rs.3.rs-5050877/v1, 2024.
- [9] R. Patil, "Real-Time Traffic Sign Detection and Recognition System using Computer Vision and Machine Learning," *Journal of Electrical Systems*, pp. 2244-2254 DOI:10.52783/jes.1991., 2024.
- [10] Z. Jefferson, "Object Detection With OpenCV: Step by Step," Medium, 21 Nov 2020. [Online]. Available: <https://medium.com/analytics-vidhya/object-detection-with-opencv-step-by-step-6c49a9cc1ff0>.
- [11] B. Sando, "PREPOZNAVANJE SAOBRAĆAJNIH ZNAKOVA PRIMJENOM RAČUNARSKOG VIDA," *Master thesis, Univerzitet Sinergija*, 2023.
- [12] K. S. Patil, S. Singh, K. V. N. A. Bhargavi, A. R. Unnithan, A. Maury and P. Dr, "Digital Image and Video Processing: Algorithms and Applications," *Journal of Electrical Systems*, p. <https://doi.org/10.52783/jes.1516>, 2024.
- [13] J. Y. Mohammed, "Enhancing The Accuracy of Image Classification Using Deep Learning and Preprocessing Methods," *Artificial Intelligence & Robotics Development Journal*, p. <https://doi.org/10.52098/airdj.2023348>, 2024.
- [14] V. Radojcic, "UPOTREBA RAČUNARSKOG VIDA ZA PREPOZNAVANJE PJEŠAKA U SAOBRAĆAJU," *Master thesis, Univerzitet Sinergija*, 2021.
- [15] V. Radojcic and M. Dobrojevic, "Transforming Traffic Safety: Detection of Car-Pedestrian Contact Using Computer Vision Technologies," *JITA - Journal of Information Technology and Applications (Banja Luka) - APEIRON*, vol. 14, no. 2, pp. 104-110 DOI: 10.7251/JIT2402104R, 2025.
- [16] V. M. B. Batta, "Image Processing using Python," *International Journal of Advanced Research in Science, Communication and Technology*, pp. <https://doi.org/10.48175/ijarsct-17499>, 2024.
- [17] A. Wedad, N. Khuder, A. K. Amal and A. S., "Subject review: Image edge detection techniques," *Global Journal of Engineering and Technology Advances*, p. <https://doi.org/10.30574/gjeta.2024.19.1.0054>, 2024.



ANT COLONY OPTIMIZATION ALGORITHM FOR FRONTEND RESOURCE PRIORITIZATION

Dejan Bulaja*,

[0009-0007-6992-979X]

Katarina Stojiljković,

[0009-0001-3718-8647]

Miodrag Živković,

[0000-0002-4351-068X]

Nebojša Bačanin Džakula,

[0000-0002-2062-924X]

Tamara Živković

[0000-0003-2969-1709]

Singidunum University,
Belgrade, Serbia

Abstract:

Optimizing frontend resource loading is crucial for enhancing web performance, as it directly affects user experience and application responsiveness. Traditional optimization techniques, such as lazy loading and dependency management, often fail to address the combinatorial complexity of resource sequencing in dynamic web environments. This paper presents a novel approach using ant colony optimization (ACO) to optimize frontend resource loading. By modeling the problem as a combinatorial optimization task, we developed a dynamic solution that considers resource dependencies, priorities, and load times. Our Python implementation demonstrates the effectiveness of ACO, achieving a 4% reduction in load time compared to the particle swarm optimization (PSO) algorithm, a 12% reduction in load time compared to the Greedy Algorithm and a 23% reduction compared to Random Loading. The algorithm consistently converges to high-quality solutions, highlighting its potential for improving web performance in complex applications. Key contributions include an ACO-based model, empirical validation, and practical insights for frontend optimization. This work underscores the value of ACO as a robust and adaptive tool for enhancing frontend performance and user experience.

Keywords:

Ant Colony Optimization, Frontend Performance Optimization, Resource Loading, Swarm Intelligence.

INTRODUCTION

Efficient frontend resource loading is a critical factor in web performance optimization [1], directly influencing user experience, search engine rankings [2], [3], and overall application responsiveness. Modern web applications rely on a complex network of interdependent resources, including HTML, CSS, JavaScript, libraries, images and other dependencies. The challenge lies in determining the optimal sequence for loading these resources to minimize delays, reduce render-blocking effects, and enhance performance. Traditional heuristics, such as predefined priority rules, fail to dynamically adapt to varying network conditions, user interactions, and device constraints, often leading to suboptimal loading times.

Correspondence:

Dejan Bulaja

e-mail:

dejan.bulaja.23@singimail.rs



Existing approaches to frontend loading optimization include techniques such as lazy loading, preloading, caching strategies, and dependency management. Lazy loading postpones resource fetching until needed, reducing initial load times, but it may introduce delays during user interaction [4]. Preloading prioritizes critical resources, but improper implementation can lead to bandwidth contention and inefficient execution order [5]. Dependency management tools analyze script relationships, but they typically operate on static configurations, lacking real-time adaptability [6]. Additionally, the process of minification [7] removes unnecessary characters from code without affecting functionality, reduces file sizes and speeds up execution, but it does not inherently address the optimal loading sequence. While these methods offer improvements in loading optimization, they remain limited in addressing the combinatorial complexity of optimizing resource sequencing in complex and dynamic web environments. Swarm Intelligence algorithms have been applied to a wide variety of practical problems [8], [9], [10], [11], [12], making them a promising approach for addressing the dynamic and combinatorial challenges of frontend loading optimization.

Ant colony optimization (ACO) [13], [14] is a biologically inspired algorithm, which offers an effective approach to the challenge of frontend resource loading. In the context of web performance, ACO can dynamically determine the best loading sequence by evaluating resource dependencies, execution priorities, and the computational cost of processing complex frontend components. By applying principles of swarm intelligence, ACO offers an adaptive, self-improving approach to minimize loading times and maximize efficiency.

This paper presents a novel ACO-based approach to optimizing frontend resource loading. Our contributions include the development of an ACO-driven model for determining the optimal loading sequence, an implementation of this model in Python, and an empirical evaluation comparing its effectiveness against conventional loading strategies. The results demonstrate the potential of ACO in reducing frontend load times, improving render efficiency, and enhancing user experience in complex web applications.

2. METHODOLOGY

2.1. PROBLEM FORMULATION

The challenge of frontend resource loading can be framed as a combinatorial optimization problem [15], where the objective is to identify the most efficient sequence for loading resources to minimize total load time while following dependency and priority constraints. Modern web applications often involve a multitude of interdependent resources, such as HTML, CSS, JavaScript, third-party libraries, and other assets, which must be orchestrated effectively to ensure optimal performance.

To model this problem, we represent the resource-loading process as a directed graph. In this graph:

- Nodes correspond to individual resources, including HTML, CSS, JavaScript, and other assets.
- Edges represent dependencies between resources, ensuring that certain elements are loaded only after their prerequisites are available.

The core challenge lies in finding an optimal path through this graph that balances execution order, minimizes blocking dependencies, and reduces unnecessary waiting times. Traditional deterministic scheduling strategies often struggle to adapt dynamically to varying page structures and execution contexts. This limitation motivates the exploration of ant colony optimization (ACO) as a promising approach to address the inherent complexity of this problem [16].

2.2. ANT COLONY OPTIMIZATION (ACO) FOR RESOURCE LOADING

ACO is a bio-inspired metaheuristic algorithm that emulates the foraging behavior of ants to solve combinatorial optimization problems. The algorithm leverages pheromone-based learning, where artificial ants explore the solution space, reinforce high-quality paths, and iteratively refine their choices.

In the context of frontend resource loading, each ant represents a potential sequence of resource loads. By iteratively constructing and evaluating different loading sequences, the system adapts dynamically to determine the best sequence that minimizes load time while ensuring that all dependency constraints are satisfied.



Key Components of ACO in our Approach

1. **Pheromone Trails:** A pheromone value is associated with each edge in the graph, representing the desirability of loading a resource in a particular order. Higher pheromone levels indicate preferable loading sequences.
2. **Heuristic Information:** Ants make decisions based on heuristic factors such as resource priority, size, and execution dependencies. This guides them toward more efficient sequences.
3. **Fitness Function:** The quality of a solution is evaluated using a fitness function that considers total load time, user engagement impact, and dependency violations.

By combining pheromone-based reinforcement with heuristic decision-making, ACO iteratively refines resource-loading sequences, ultimately converging toward an optimal or near-optimal solution.

2.3. ALGORITHM IMPLEMENTATION

Our ACO-based algorithm for frontend optimization follows the structured workflow outlined below.

1. Initialization

The initialization phase begins by setting pheromone levels uniformly across all edges, ensuring a balanced starting point for exploration. Next, heuristic functions are defined based on resource priority, execution complexity, and the expected impact on load performance to guide the solution search effectively. Finally, key ACO parameters such as the number of ants, pheromone evaporation rate, and the balance between exploration and exploitation are configured to optimize convergence and solution quality.

2. Solution Construction

During the solution construction phase, each ant sequentially selects resources to build a candidate loading sequence while ensuring that all dependency constraints are satisfied. The selection process is influenced by a pheromone-weighted probability rule, where resources with higher pheromone levels and stronger heuristic desirability are more likely to be chosen. This balance between learned experience (pheromones) and heuristic guidance allows the system to explore various loading sequences while favoring those that are more efficient. Through repeated iterations, ants collectively refine their choices, progressively identifying optimal or near-optimal solutions. The probability of selecting a resource at each step is governed by a pheromone-weighted selection rule:

$$P_{i,j} = \frac{\tau_{i,j}^{\alpha} \cdot \eta_{i,j}^{\beta}}{\sum_{k \in \text{available}} \tau_{i,k}^{\alpha} \cdot \eta_{i,k}^{\beta}}$$

Where:

$\tau_{i,j}$: The pheromone level on the edge from resource i to resource j

$\eta_{i,j}$: Heuristic desirability of selecting resource j after resource i , influenced by priority and size

α and β control the balance between pheromone influence and heuristic information.

3. Pheromone Update

After all ants construct their sequences, pheromone levels are updated based on solution quality:

$$\tau_{i,j} = (1 - \rho)\tau_{i,j} + \sum_k^m \Delta\tau_{i,j}^k$$

Where:

ρ : The evaporation rate, preventing premature convergence.

$\Delta\tau_{i,j}^k$: The contribution of ant k , with stronger solutions depositing more pheromones.

4. Termination Criteria

The process iterates for a predefined number of cycles or until convergence, meaning the solution quality stabilizes with minimal further improvement. As the algorithm progresses, pheromone updates guide the ants toward increasingly optimal sequences. Once the termination condition is met, the best-performing resource sequence which represents the most efficient frontend loading plan is selected and returned as the final optimized solution.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The experiment in this paper is designed to explore the potential of ACO for frontend resource loading optimization.

3.1. EXPERIMENTAL SETUP

The experiments were conducted in a simulated web environment using Google Chrome (version 133.0) on a MacOS system (version 15.3). The algorithm is implemented in Python, using libraries like numpy for numerical computations, math for providing access to the mathematical functions, random for generating pseudo-random numbers and matplotlib for visualizations. The choice of parameters, such as the number of ants, evapo-



ration rate (0.3), and dependency penalty weights (α and β), was informed by empirical tuning and established practices in ACO literature. Gaertner and Clark [17] demonstrated that optimal parameter settings are highly problem-specific and depend on the required solution accuracy. Their findings suggest that a higher evaporation rate and balanced exploration-exploitation often yield robust performance across diverse problem instances. Based on their insights, we calibrated our parameters through iterative testing to ensure a balance between load time optimization and dependency constraint satisfaction.

The fitness function is a critical component of the ACO implementation, as it evaluates the quality of a solution. Our proposed fitness function is:

$$\text{Fitness} = \frac{\sum_{i=1}^n \omega_i(t) \cdot \text{Priority}_i}{\log_2(\text{Load Time} + 1)} - \alpha \cdot \text{Dependency Penalty} - \beta \cdot \text{Size Penalty}$$

Where:

$\omega_i(t)$: Dynamic user engagement weight for resource i at time t

Priority_i : Priority of resource i

Load Time : Total load time

$\text{Dependency Penalty}$: Penalty for dependency violations

Size Penalty : Penalty for loading large resources early

α, β : Tuning parameters to balance the penalties

We simulate a frontend environment with the following resources:

1. HTML: The base structure of the web page, containing the content and layout.
2. CSS: Styling resources that define the appearance of the web page (e.g., colors, fonts, layout).
3. JavaScript: Interactive elements and functionality (e.g., animations, form validation).
4. Framework: A JavaScript or CSS framework (e.g., React, Bootstrap) that provides reusable components and utilities
5. Fonts: Custom fonts used for text rendering on the web page
6. Icons: Small graphical elements (e.g., social media icons, navigation icons)
7. Analytics: Tracking scripts (e.g., Google Analytics) that collect user behavior data.
8. Images: Visual content (e.g., photos, illustrations) displayed on the web page.
9. Ads: Advertisement scripts and content displayed on the web page.
10. Videos: Embedded video content (e.g., YouTube videos, promotional videos).

The user engagement weights are derived from pre-defined priorities assigned to each resource, reflecting their relative importance to the user experience. For example, critical resources like HTML and CSS are assigned higher priorities (5 and 4, respectively), while less critical resources like ads and videos receive lower priorities (1). These priorities are based on the resource's role in the rendering pipeline and its impact on initial page load performance, ensuring that the ACO algorithm optimizes for resources that most directly influence user engagement. The weights are further informed by dependencies between resources, such as JavaScript requiring HTML to load first, ensuring that the loading sequence aligns with both performance goals and functional requirements. This approach ensures a systematic and reproducible basis for resource prioritization.

3.2. EXPERIMENTAL RESULTS AND DISCUSSION

We conducted experiments to evaluate the performance of the ACO algorithm in optimizing frontend resource loading. The algorithm was run for 30 iterations with 20 ants, and the results were analyzed using the following metrics:

- Fitness: A combination of user engagement and load time, as defined by our fitness function
- Load time: The total time required to load all resources

The ACO algorithm achieved the best fitness of 0.981. The best resource loading sequence found by the algorithm is shown in Figure 1.

In Figure 1, the best resource loading sequence found by the ACO algorithm is illustrated, demonstrating the optimized order of resource loading that resulted in a total load time of 1200 ms, with fitness values improving over iterations, as shown in Figure 2.

Figure 2 shows the convergence of fitness values over iterations, highlighting the consistent improvement in solution quality as the ACO algorithm progresses toward an optimal or near-optimal resource loading sequence.

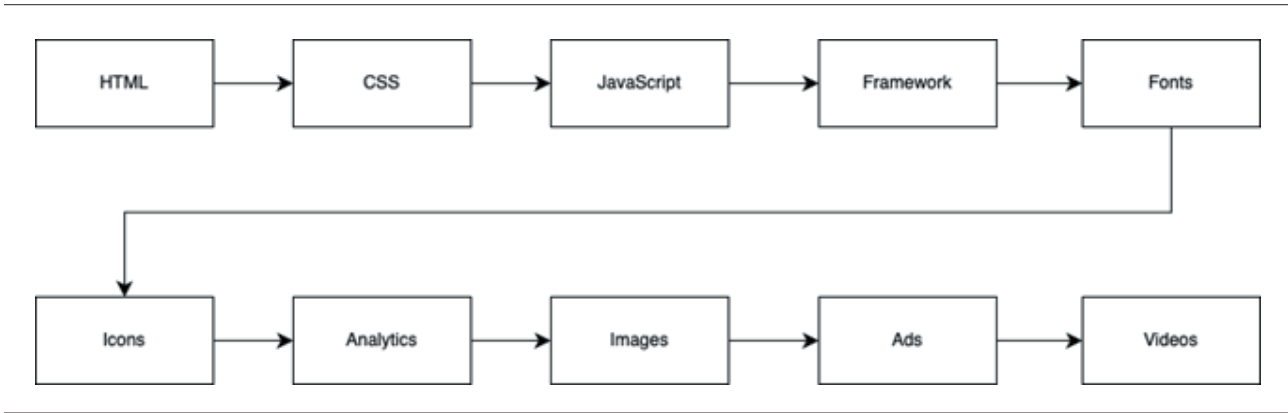


Figure 1. Best resource loading sequence

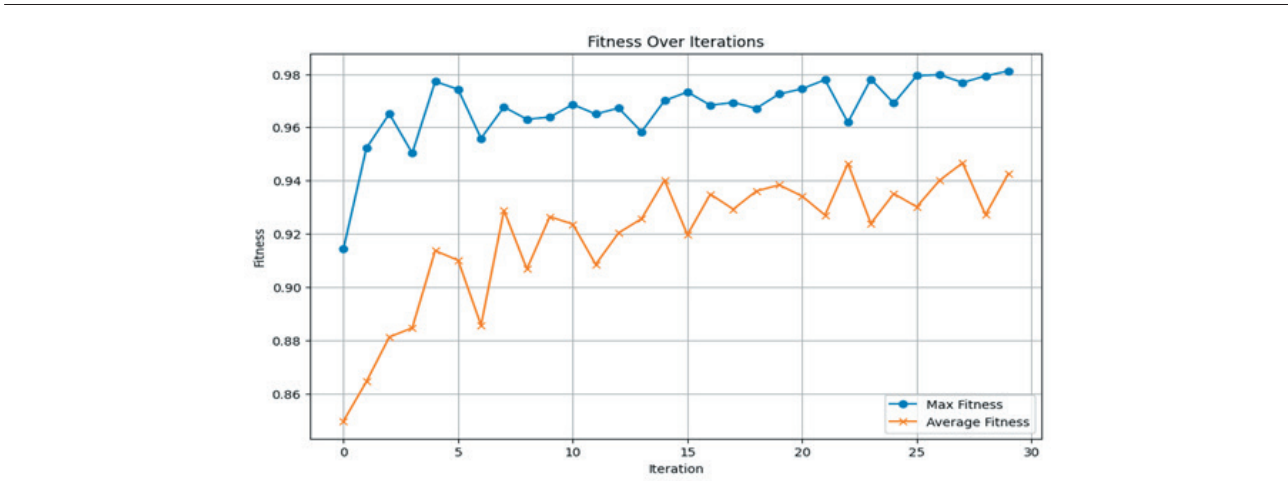


Figure 2. Fitness over iterations

3.3. COMPARISON WITH BASELINES

We compared the ACO algorithm with two baseline approaches: Random Loading and the Greedy Algorithm. Additionally, to provide a more comprehensive evaluation, we implemented a Particle Swarm Optimization (PSO) algorithm [18], a modern metaheuristic optimization technique, and compared its performance with ACO.

1. Random Loading: Resources are loaded in a random order while respecting dependencies
2. Greedy Algorithm: Resources are loaded in order of their priority (higher priority first), breaking ties by load time (shorter load time first).
3. Particle Swarm Optimization (PSO): A population-based optimization algorithm where particles (solutions) move through the solution space to find optimal resource loading sequences. The algorithm balances exploration and exploitation by adjusting particle velocities based on their own best-known position and the global best-known position.

The results of the comparison are summarized below:

Algorithm	Best Fitness	Average Load Time (ms)
ACO	0.981	1210
Greedy Algorithm	0.967	1380
Random Loading	0.942	1570
PSO	0.975	1260

The ACO algorithm outperformed all baselines, including PSO, in terms of fitness and load time. While PSO achieved a competitive fitness of 0.975 and an average load time of 1260 ms, ACO demonstrated superior convergence and efficiency, achieving a 12% reduction in load time compared to the Greedy Algorithm, a 23% reduction compared to Random Loading, and a 4% reduction in load time compared to the PSO algorithm.



The findings suggest that ACO can be a valuable tool for frontend developers seeking to improve web performance. By considering dependencies, priorities, and load times, the algorithm provides a systematic way to optimize resource loading sequences. This is particularly useful for complex web applications with many interdependent resources.

4. CONCLUSION

This paper introduced a novel ant colony optimization (ACO) approach for optimizing frontend resource loading. By modeling the problem as a combinatorial optimization task, we developed a dynamic solution that considers dependencies, priorities, and load times. Our Python implementation demonstrated ACO's effectiveness, achieving a 4% reduction in load time compared to the PSO algorithm, a 12% reduction in load time compared to the Greedy Algorithm and a 23% reduction compared to Random Loading. The algorithm consistently converged to high-quality solutions, highlighting its potential for improving web performance in complex environments.

The key contributions of this work include the development of an ACO-based model for frontend resource loading, designed to enhance web performance. Through empirical validation, the proposed approach demonstrates superior performance compared to baseline methods. Additionally, the findings offer practical implications for optimizing resource sequencing in modern web applications, improving efficiency and user experience.

The current approach faces limitations, including scalability challenges with large resource sets and a reliance on predefined dependencies. To address these issues and enhance the system's capabilities, future work could explore several promising directions. One way is dynamic resource loading, where the algorithm adapts to handle on-demand resource loading based on user interactions, improving efficiency and responsiveness. Additionally, real-world deployment on live websites with actual user data would validate the approach's effectiveness in practical settings, ensuring its applicability beyond theoretical scenarios. Furthermore, integrating machine learning could enable the prediction of user engagement weights, allowing for dynamic adjustments to resource priorities and fostering adaptive optimization. These advancements would collectively address existing limitations and unlock new potential for the system.

REFERENCES

- [1] B. Weischedel and E. Huizingh, "Website optimization with Web metrics: A case study," in *Proceedings of the ACM Conference on Electronic Commerce*, Feb. 2006, pp. 463–470. doi: 10.1145/1151454.1151525.
- [2] D. Y. Pawade, "Analyzing the impact of search engine optimization techniques on web development using experiential and collaborative learning techniques," *International Journal of Modern Education and Computer Science*, vol. 13, no. 2, pp. 1–10, 2021, doi: 10.5815/IJMECS.2021.02.01.
- [3] M. Cui and S. Hu, "Search Engine Optimization Research for Website Promotion," *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, vol. 4, pp. 100–103, 2011, [Online]. Available: <https://api.semanticscholar.org/CorpusID:36037694>
- [4] A. Turcotte, S. Gokhale, and F. Tip, "Increasing the Responsiveness of Web Applications by Introducing Lazy Loading," in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2023, pp. 459–470. doi: 10.1109/ASE56229.2023.00192.
- [5] L. Cegan, "Intelligent Preloading of Websites Resources Based on Clustering Web User Sessions," in *2015 5th International Conference on IT Convergence and Security (ICITCS)*, 2015, pp. 1–4. doi: 10.1109/ICITCS.2015.7292988.
- [6] I. Pashchenko, D.-L. Vu, and F. Massacci, "A Qualitative Study of Dependency Management and Its Security Implications," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, in CCS '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1513–1531. doi: 10.1145/3372297.3417232.
- [7] Y. Sakamoto, S. Matsumoto, S. Tokunaga, S. Saiki, and M. Nakamura, "Empirical study on effects of script minification and HTTP compression for traffic reduction," *2015 3rd International Conference on Digital Information, Networking, and Wireless Communications, DINWC 2015*, pp. 127–132, Feb. 2015, doi: 10.1109/DINWC.2015.7054230.
- [8] M. Dobrojevic *et al.*, "Cyberbullying Sexism Harassment Identification by Metaheuristics-Tuned eXtreme Gradient Boosting," *Computers, Materials & Continua*, vol. 80, no. 3, pp. 4997–5027, 2024, doi: 10.32604/cmc.2024.054459.
- [9] D. Mladenovic *et al.*, "Sentiment classification for insider threat identification using metaheuristic optimized machine learning classifiers," *Sci Rep*, vol. 14, no. 1, p. 25731, 2024, doi: 10.1038/s41598-024-77240-w.



- [10] A. Stojkovic, B. Nikolic, M. Zivkovic, and N. Bacanin, "Photovoltaic Farm Production Forecasting: Modified Metaheuristic Optimized Long Short-Term Memory Based Networks Approach," *IEEE Access*, Feb. 2025.
- [11] F. Markovic et al., "Parkinsons Detection from Gait Time Series Classification Using Modified Metaheuristic Optimized Long Short Term Memory," *Neural Process Lett*, vol. 57, no. 1, p. 14, 2025, doi: 10.1007/s11063-025-11735-z.
- [12] P. Mani and N. Bacanin, "Employee reviews sentiment classification using BERT encoding and Ada-Boost classifier tuned by modified PSO algorithm," in *Proceedings of the 2nd International Conference on Innovation in Information Technology and Business (ICIITB 2024)*, 2024, p. 22.
- [13] M. Dorigo, M. Birattari, and T. Stützle, "Ant Colony Optimization," *Computational Intelligence Magazine, IEEE*, vol. 1, pp. 28–39, Feb. 2006, doi: 10.1109/MCI.2006.329691.
- [14] T. Dorigo Marco and Stützle, "Ant Colony Optimization: Overview and Recent Advances," in *Handbook of Metaheuristics*, J.-Y. Gendreau Michel and Potvin, Ed., Cham: Springer International Publishing, 2019, pp. 311–351. doi: 10.1007/978-3-319-91086-4_10.
- [15] F. Peres and M. Castelli, "Combinatorial optimization problems and metaheuristics: Review, challenges, design, and development," *Applied Sciences (Switzerland)*, vol. 11, no. 14, Jul. 2021, doi: 10.3390/app11146449.
- [16] K. M. Sim and W. H. Sun, "Ant colony optimization for routing and load-balancing: survey and new directions," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 33, no. 5, pp. 560–572, 2003, doi: 10.1109/TSM-CA.2003.817391.
- [17] K. Clark and D. Gaertner, "On Optimal Parameters for Ant Colony Optimization Algorithms. On Optimal Parameters for Ant Colony Optimization algorithms," 2005. [Online]. Available: <https://www.researchgate.net/publication/220835272>
- [18] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1995, pp. 1942–1948 vol.4. doi: 10.1109/ICNN.1995.488968.



CYBERSECURITY THREATS FOR MEDICAL IMAGING DEVICES: A SYSTEMATIC REVIEWE

Shakeel Ahmed^{1*},
[0009-0001-8508-819X]

Samina Khalid¹,
[0000-0003-4771-6842]

Yasir Mehmood¹,
[0000-0003-3074-8944]

Modestus O. Okwu²
[0000-0002-7761-9659]

¹Mirpur University of Science and
Technology,
Mirpur, Azad Kashmir, Pakistan

²Federal University of Petroleum
Resources Effurun,
Warri, Nigeria

Abstract:

The innovation and modernization in healthcare infrastructure have been achieved by employing IT infrastructure in public healthcare setups ranging from Hospital Information System (HIS) to all the digital gadgets which are helping medical facilities to achieve more productivity. The use of digital medical imaging devices in advanced diagnostic systems has enhanced diagnostic precision and refined disease treatment. On the other hand, these developments also introduced a new challenge of cybersecurity issues, such as data breaches, unauthorized access, and ransomware attacks on healthcare facilities. A systematic literature review was performed across prominent databases, concentrating on recent developments in cybersecurity for medical imaging. Despite growing concerns, a comprehensive analysis of cybersecurity risks and countermeasures specific to medical imaging devices remains scarce. This review aims to bridge this gap by systematically analyzing existing threats, vulnerabilities, and mitigation strategies. The potential risks, precautionary measures, protocols, and probable mitigation strategies are highlighted in detail.

Keywords:

Medical Imaging, Digital Imaging and Communications in Medicine (DICOM), Picture Archiving and Communication System (PACS), Cyber Security, Artificial Intelligence.

INTRODUCTION

Medical imaging devices are increasingly vulnerable to cyber threats like data breaches, image manipulation, and ransomware, putting patient safety at risk. With healthcare relying more on AI-driven diagnostics, securing these systems is crucial to prevent misdiagnoses, protect patient privacy, and maintain trust in medical technology. The advent of medical imaging technologies like Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scan and X-Ray help medical practitioners to diagnose problems in patients in a more timely and effective manner. Since their advent and employment in the medical field the potential security concerns have also increased. Today, medical imaging devices are under serious threats by attackers who with their criminal instincts try to hack into hospital networks.

Correspondence:

Shakeel Ahmed

e-mail:

shakeel803@gmail.com





Digital medical imaging technologies, including MRI, CT, and X-ray apparatus, have transformed contemporary healthcare by facilitating accurate diagnosis and treatment strategies. Nonetheless, their escalating connectivity with hospital networks, cloud storage, and AI-driven diagnostics has rendered them great targets for assaults. Ransomware assaults represent a significant concern, as malicious software encrypts image data, making it inaccessible until a ransom is remitted. Such assaults can impede hospital operations, postpone critical medical treatments, and jeopardize patient safety. Despite the increasing prevalence of ransomware events in healthcare, the particular vulnerabilities of medical imaging devices continue to be an inadequately examined domain in cybersecurity research. This study offers an extensive analysis of ransomware attacks aimed at imaging systems, examines their attack methodologies, and evaluates novel defense techniques to maintain data integrity and ensure continuity of patient care.

Although current research addresses cybersecurity issues in healthcare, the majority concentrates on general hospital networks or electronic health records, thereby neglecting medical imaging devices. The growing use of AI and cloud storage in imaging systems creates new risks, like attacks on diagnostic models and ransomware targeting these devices. However, there is a lack of comprehensive reviews that systematically analyze these specific risks, their real-world implications, and potential mitigation strategies. This review aims to bridge this gap by providing an in-depth analysis of cybersecurity threats unique to medical imaging devices and proposing future research directions to enhance their security.

A document released by the US Department of Justice indicates that 4,000 ransomware attacks have been recorded daily since 2016[1] which is four times increase as compared to 2015. Out of all global ransomware attacks, 15% were attacking healthcare facilities in 2017. As far as the cyber security related to medical imaging devices is concerned, the primary and possibly most evident aspect of cybersecurity is physical security; technical mitigation strategies such as passwords, antivirus software, or detailed user permissions are ineffective if an intruder can easily access a server room and abscond with computers or storage devices [1].

DICOM is a standard for storing and transmitting medical images and related information. DICOM standards are developed by The American College of Radiology (ACR) and The National Electrical Manufacturers Association (NEMA) in 1985. DICOM governs the standards to store and transfer data of medical imaging in healthcare facilities.

PACS is a medical imaging technology that stores, retrieves, and transmits digital medical images and reports. PACS systems are used in healthcare organizations to replace the need for film jackets by electronically storing and transmitting images and reports.

While we talk about the cyber security threats related to digital imaging in the medical field, we are concerned about all the aspects that can pose a risk to the DICOM or PACS based infra structure of any healthcare facility. Cybersecurity threats have been increasing as hospitals become more attractive targets for cybercriminals. This is because healthcare organizations are often willing to pay higher ransoms to protect their reputation and maintain the trust of their patients and stakeholders.

Safeguarding healthcare facilities is of paramount importance so as to safeguard the personal patient data as well as the IT infrastructure of healthcare facilities. Hospitals within the United Kingdom's National Health Service, impacted by the WannaCry ransomware attacks in May 2017, were compelled to postpone treatment plans and redirect arriving ambulances due to the loss of access to hospital information systems [2]. Cyberattacks pose a significant risk to numerous hospital services, including surgeries and medication distribution, by compromising sophisticated devices such as blood-product refrigerators, imaging apparatus, automated drug dispensers, and electronic health records, in addition to essential support systems like heating, ventilation, and air conditioning (HVAC) [2]. An individual's health information is considerably more valuable on the dark web than their social security number or credit card number, perhaps selling for 10 to 20 times more than such data [2].

The healthcare sector is increasingly becoming a prime target for cyber-attacks. A survey of 223 organizations revealed that 81% were impacted by cyber-attacks, and more than 110 million patients in the USA had their data exposed in 2015 [3]. The utilization of end-of-life software presents a heightened danger, as evidenced by assaults on healthcare providers, like the Irish Health Service Executive's "Conti" and the UK National Health Service's "WannaCry" ransomware incidents [3].

Optimal cybersecurity necessitates a collaborative approach, making it essential for the people working in healthcare field to remain cognizant of their responsibilities in individual as well as group regarding the security of healthcare facility they work for. Storage and transmission of various confidential data elements, including medical history, dates of birth and financial information is mandatory in radiological diagnosis and



other healthcare procedures. Consequently, healthcare organizations face monetary, legal, and reputational risks if information security is breached [4]. Healthcare data breaches in the US involving 500 or more patients are reported by the responsible organizations publicly and an upward trend has been witnessed since 2014. Although, the incidents of such breaches in radiology and digital imaging procedures are in low fraction, yet there has been an upward trend [4]. The medical imaging environments are implementing multiple techniques including De-identification of medical images, securing DICOM and PACS transmission, encrypting the image data completing before transferring, digital signature, watermarking etc [1].

One of the biggest challenges for any organization is the lack of IT knowledge among the medical workforce. In the study [5], the author has discussed in detail all the must to know factors of the DICOM infrastructure and has also proposed a simulation-based approach to implement, study and diagnose potential risks in a DICOM-based network setting.

This review paper is further divided in sections. In Section 2, general cyber security considerations will be discussed in detail. In section 3 the cyber security risks related to the DICOM and PACS based medical imaging infrastructure will be discussed. Section 4 is a discussion on available options to safeguard the digital imaging devices and lastly the conclusion of the paper is discussed in section 5.

2. CYBER SECURITY: GENERAL CONSIDERATIONS

Cyber security is one of the emerging fields of computer science which deals with safety of computer systems against the criminal intents. One of the common cyber security threats is DoS (Denial of Service) attack in which attackers over flood any organization's IT infrastructure with unreal requests and the system could slow down and the genuine users may face access problems. Similarly, the attacks to steal data and sell it on dark web is another common aspect of cyber security. In the analysis of various industries, DoS attacks represent the predominant form of cybersecurity incidents, accounting for over 50%. Although these attacks are disruptive, they generally do not lead to unauthorized access to data [4].

In addition to DoS and malware, various other mechanisms may compromise data security. Privilege escalation is the enhancement of access rights of a legitimate user, such as increasing the access to an adminis-

trator-level account is very dangerous as it enables a user to install and execute malicious software. Eavesdropping attacks involve the unauthorized observation or alteration of information shared between two devices by masquerading as an intermediary, while cryptographic attacks aim to uncover hidden data. Significant volumes of data are exchanged and saved across multiple electronic systems daily in X-Ray and radiology departments of healthcare facilities. Privacy of patients and confidentiality are exposed to high risks due to such attacks, potentially compromising sensitive institutional data or proprietary information [4].

3. CYBER SECURITY CHALLENGES TO MEDICAL IMAGING DEVICES

3.1. IMPORT OF PATIENT DATA

A malware infected storage media used for transferring medical images can import all patient data when used with the healthcare's IT infrastructure to conduct the study. In many scenarios, the patient can bring along a storage media in which the studies of their scans (conducted by another entity) are present that need to be evaluated. In this case, the data available on the hospital network could be breached. To overcome this, the viewers used by the 3rd party must not be allowed to autorun rather own viewers be used to view the studies. Most systems that create DICOM CDs also write an executable viewer that could be a potential risk as well.

3.2. HACKING OF THE HOSPITAL NETWORK

The infiltration of hospital networks presents a substantial and escalating risk to patient safety, data confidentiality, and the entire operation of healthcare systems. As hospitals increasingly depend on interconnected technologies, including medical imaging apparatus, Electronic Health Records (EHR), and other essential infrastructure, cyberattacks such as ransomware, data breaches, and system intrusions are evolving in sophistication and severity. These breaches jeopardize patient care and disrupt hospital operations, resulting in financial losses and regulatory penalties. Healthcare institutions must have comprehensive cybersecurity frameworks that encompass robust network segmentation, regular system updates, stringent access control measures, and continuous monitoring to safeguard against the constantly expanding cyber threat scenario. Securing hospital networks is imperative; it is essential for preserving trust in healthcare systems and safeguarding



the security and privacy of sensitive patient information. This is the worst situation in which cyber criminals get access to the hospital's LAN using an unprotected port or compromising a wireless device. Once the hacker has access to the network, he could passively monitor all the traffic over the network. The data transferred over DICOM and HL7 version 2 is in plain text format and the hacker can use some packet analyzer to sniff all the data and gather all the information about images, patient data and the network addressing of the DICOM servers. When the attacker knows address of all the servers, he could easily gain unauthorized access to servers and can manipulate highly sensitive and confidential data.

3.3. MANIPULATION OF MEDICAL IMAGES WITH A MALICIOUS INTENT

With the advent and progression of Artificial Intelligence (AI), the image manipulation techniques have evolved at a skyrocketing speed. Deep Fake can be used to corrupt the medical images of patients and expert radiologists may remain unable to sense the forged and actual images. This forgery can be life threatening as the diagnosis may go astray. Steganography is used to protect original files so that the data can't be manipulated by the hackers. Steganography pertains to the concealing of confidential information, encompassing its storage and transmission. A variety of digital artifacts have been examined concerning both steganography application and its detection, including digital text, photos, videos, audio files, filesystems, cyber-physical systems, and networks [6].

Mirsky et al. state that "to verify the threat of this attack, we trained CT-GAN to inject/remove lung cancer and hired three radiologists to diagnose a mix of 70 tampered and 30 authentic CT scans. The experiment was performed in two trials: blind and open. In the blind trial, the radiologists were asked to diagnose 80 complete CT scans of lungs, but they were not told the purpose of the experiment or that some of the scans were manipulated. In the open trial, the radiologists were told about the attack, and were asked to identify fake, real, and removed nodules in 20 CT scans. In addition, the radiologists were asked to rate the confidence of their decisions." The results were quite impressive: in the blind trial, "the radiologists diagnosed 99% of the injected patients with malignant cancer, and 94% of cancer removed patients as being healthy. After informing the radiologists of the attack, they still misdiagnosed 60% of those with injections, and 87% of those with removals [7]".

4. CYBER SECURITY PROTOCOLS TO MITIGATE THE RISKS

4.1. SEGMENTATION OF THE NETWORK

It is a process of distributing the bigger networks into smaller easy to manage and troubleshoot networks. The network segmentations help in safeguarding the complete network of healthcare facilities when one segment is compromised. We just need to shut down that specific segment and keeping rest of the healthcare IT infrastructure.

Local area network (LAN) could be laid down by a college or university to restrict access to on-campus computers and to isolate it from public internet. On similar lines, devices installed at homes are maintained on a local area network and separation is achieved from the internet. Network segmentation and segregation enable enterprises to enhance security standards and/or restrict access to devices with elevated security requirements, thereby isolating essential segments of network traffic from non-essential traffic Enterprises use Network segmentation and segregation to improve security protocols and to restrict unauthorized access to the devices which need more confidentiality. Using this technique organizations segregate their essential and non-essential network traffic [4].

4.2. REGULAR BACKUPS

ENISA (European Union Agency for Network and Information Security) recommends the performance of regular backups. "This very important action can solve many attacks that could cause great impacts to smart hospitals such as ransomware or physical attacks. Running regular full or incremental backups can be done combined with setting a hot or warm site, making the hospital systems resilient even in the case of natural disaster [1]." Although regular backups do not act as an active measure to safeguard the network, with the help of regular backups, we could ensure the availability of the most recent data in case of the data breach or theft. This will ensure that healthcare facility is having most recent data with them somewhere at a cold backup storage facility from where the loss could be mitigated in case of any unforeseen scenarios.



4.3. USE OF ENCRYPTION TECHNIQUES

Various encryption techniques are being used to safeguard the personal identity information of patients while transferring data using DICOM and PACS infrastructures. Extensive efforts have been undertaken to protect the data during transmission to ensure its integrity is preserved. One of the developed techniques is the application of steganographic methods to the DICOM Message Service and Upper Layer Service to establish hidden channels [5]. A simple XOR based encryption technique has been introduced in which reduced the encryption decryption times for multi-frame DICOM-based images [8].

One of the major challenges of encryption is that huge amount of computing is required to encrypt and decrypt the heavy DICOM images. In a study [9], a lightweight two permutation-based algorithm to encrypt medical images is proposed. The security and execution time of the suggested method are examined, assessed, and then contrasted with those of traditional encrypted methods. The performance of the suggested approach has been evaluated using a large number of test photos. Numerous tests demonstrate that the suggested approach outperforms traditional methods in terms of efficiency for picture cryptosystems.

4.4. AUTHENTICATION AND ACCESS MANAGEMENT

Access permissions and user authentication procedures serve as a crucial barrier to unauthorized system access. A prevalent cybersecurity measure is to refrain from utilizing accounts with administrative capabilities unless explicitly necessary for actions such as software installation or updates. This access significantly increases cybersecurity risks compared to accounts with restricted access, and the concept of least privilege must be adhered to when providing users access to electronic resources. Protocols for minimum password strength and the frequency of password alterations can be established to protect against unwanted access to the DICOM and PACS systems. Latest security trends such as multi-factor authentication, role-based access control, and Zero Trust security models must be adopted to safeguard the DICOM and PACS systems installed at healthcare facilities. If we overlook these aspects and keep the system open to all without proper access rights, we are inviting the wrong doers to attack and gain the access to the healthcare network.

4.5. DISABLING UNUSED PHYSICAL PORTS

If the PACS server is having unused open ports, there is a high probability that a criminal can get advantage of it to sneak into the sensitive information of patients using techniques. An impeccable defense against a hacker infiltrating the hospital network will be of minimal utility if the PACS server is entirely accessible and responds to inquiries from across the Internet. As most of the end users of Digital Medical Imaging devices are people who do not have knowledge of the cyber security practices, we can't blame those people. The IT teams employed at healthcare facilities must strive hard to ensure that deployed systems are not prone to such attacks where the unused ports could be manipulated to gain access to hospital networks. This is, unfortunately, not a theoretical issue. A 2019 investigation revealed that numerous PACS systems worldwide are vulnerable to Internet access due to an absence of fundamental IT security protocols [1].

5. CONCLUSION

This systematic review sought to examine the emerging cybersecurity threats to medical imaging devices and evaluate current vulnerabilities, possible dangers, and mitigation options within healthcare systems. As the integration of digital technologies such as DICOM and PACS becomes more widespread, these systems have become prime targets for cyberattacks—including ransomware, unauthorized access, and image manipulation—which can severely impact patient safety and healthcare operations. Our findings indicated that numerous healthcare facilities continue to function with outdated systems, unsecured network ports, and insufficient user training, rendering them progressively vulnerable to both internal and external attacks. The review emphasizes various effective defense systems, including network segmentation, routine system backups, encryption methods, and stringent authentication and access control procedures. Moreover, forthcoming initiatives should concentrate on closing the knowledge gap among healthcare personnel, using AI-based security solutions, and enforcing uniform cybersecurity protocols. Enhancing these domains is crucial for safeguarding sensitive medical information, ensuring system accessibility, and maintaining trust in the digital evolution of healthcare.



REFERENCES

- [1] M. Eichelberg, K. Kleber, and M. Kämmerer, "Cybersecurity in PACS and Medical Imaging: an Overview," *J. Digit. Imaging*, vol. 33, no. 6, pp. 1527–1542, Dec. 2020, doi: 10.1007/s10278-020-00393-3.
- [2] S. T. Argaw et al., "Cybersecurity of Hospitals: discussing the challenges and working towards mitigating the risks," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 146, Dec. 2020, doi: 10.1186/s12911-020-01161-7.
- [3] B. S. Kelly, C. Quinn, N. Belton, A. Lawlor, R. P. Killeen, and J. Burrell, "Cybersecurity considerations for radiology departments involved with artificial intelligence," *Eur. Radiol.*, vol. 33, no. 12, pp. 8833–8841, Jul. 2023, doi: 10.1007/s00330-023-09860-1.
- [4] X. V. Nguyen, J. M. Petscavage-Thomas, C. M. Straus, and I. Ikuta, "Cybersecurity in radiology: Cautionary Tales, Proactive Prevention, and What to do When You Get Hacked," *Curr. Probl. Diagn. Radiol.*, p. S0363018824001221, Jul. 2024, doi: 10.1067/j.cpradiol.2024.07.010.
- [5] S. Karagiannis, E. Magkos, C. Ntantogian, R. Cabecinha, and T. Fotis, "Cybersecurity and Medical Imaging: A Simulation-Based Approach to DICOM Communication," *Appl. Sci.*, vol. 13, no. 18, p. 10072, Sep. 2023, doi: 10.3390/app131810072.
- [6] A. Mileva, A. Velinov, V. Dimitrova, L. Cavigliione, and S. Wendzel, "Information Hiding in the DICOM Message Service and Upper Layer Service with Entropy-Based Detection," *Entropy*, vol. 24, no. 2, p. 176, Jan. 2022, doi: 10.3390/e24020176.
- [7] M. Eichelberg, K. Kleber, and M. Kämmerer, "Cybersecurity Challenges for PACS and Medical Imaging," *Acad. Radiol.*, vol. 27, no. 8, pp. 1126–1139, Aug. 2020, doi: 10.1016/j.acra.2020.03.026.
- [8] Q. N. Natsheh, B. Li, and A. G. Gale, "Security of Multi-frame DICOM Images Using XOR Encryption Approach," *Procedia Comput. Sci.*, vol. 90, pp. 175–181, 2016, doi: 10.1016/j.procs.2016.07.018.
- [9] M. K. Hasan et al., "Lightweight Encryption Technique to Enhance Medical Image Security on Internet of Medical Things Applications," *IEEE Access*, vol. 9, pp. 47731–47742, 2021, doi: 10.1109/ACCESS.2021.3061710.



THE IDS SECURITY CHALLENGE SOLUTIONS OFFERED BY METAHEURISTIC OPTIMIZATION

Dušan Cvetković *,
[0009-0006-2436-4740]

Miodrag Živković,
[0000-0002-4351-068X]

Nebojša Bačanin Džakula
[0000-0002-2062-924X]

Singidunum University,
Belgrade, Serbia

Abstract:

The concern of this paper is to analyze the potential solutions offered by metaheuristic optimization for intruder detection systems, which have become standard due to their use throughout industries. Most recent trends have caused a large influx of potentially vulnerable devices, leading to the increasing challenge of properly detecting what constitutes a false positive or true positive detection. The aim of the research is twofold:

- a) to identify aspects of the intrusion detection system that can be improved
- b) to identify methods via which this improvement can be achieved.

The methodology of meta-research includes a comparative analysis of the systems based on secondary sources (papers published in prestigious journals) and accompanying references to the theoretical and industrial aspects.

The first step is to analyze the optimization techniques, chosen as the case studies, such as the genetic algorithm, firefly algorithm, chimp optimization algorithm, etc. In the next step, the paper diagnoses the security challenges faced by modern IDS solutions and discusses the proposed improvement (and optimizations) offered by the previously mentioned metaheuristic optimizations.

Keywords:

Intrusion Detection System, Classification Method, Metaheuristic Optimization, Meta-analysis.

INTRODUCTION

IDS is one of the key constituents of a comprehensive cybersecurity infrastructure. It is used for monitoring various communications, both within the confines of an organization's network and outside it, as out-bound traffic. One of the most vital roles it performs is the detection of anomalous activity, regardless of whether it is a new unknown address attempting to access a server secured only for privileged users or it is a sudden large transfer of files to an unknown destination (device, server, or IP address). With a steep increase in the number of devices, the average user employs due to the development of IoT, systems like IDS, when it comes to parsing which alerts and detections are genuine and which are not, are faced with a great challenge.

Correspondence:

Dušan Cvetković

e-mail:

dusan.cvetkovic.24@singimail.rs





Many attacks remain invisible to an automated system because they inhabit the border between the rational and emotional. In these cases, the strategy of AI deployment – regarding which algorithm would be best suited to uncover such attacks – must be considered in advance. The inherent cost-benefit analysis reveals that running an AI model is still expensive, especially if it is to be dedicated to a singular purpose. The cost can be both financial and technical, in terms of limited processing power available to a machine. When considering the different kinds of machine learning models, the classification models stand out as the most optimal solutions. Classification machine learning models focus on correctly labeling and identifying sets of data (a good example would be the separation of spam emails from legitimate correspondence).

Primarily, the model must be able to separate the intentional and unintentional insider malicious actors, [1] – the latter being users who become malicious actors by accident. When security policies and rules are not followed properly, they create an opening for malicious actors to exploit. Early detection of such occurrences is a way of improving the onboarding and security education of users [2]. On the opposite side, there are intentional malicious actors, the masquerader and the traitor. Masqueraders acquire access to the system through existing compromised credentials. Through impersonation, they seek to gain access to sensitive materials and vulnerable systems before abandoning their identity, and it is not uncommon for a user's unintentional action to create a potential opening for an attack by a masquerader. The most dangerous threat is the traitor – a legitimate user who leverages his access, knowledge and trust to further malicious activity within an organization [3]. A machine learning model can be based on different learning styles, specifically bagging and boosting. In this paper, the focus is on boosting learning models characterized by the sequential processing of data.

2. CHALLENGES FACED BY IDS

The majority of current-day intrusion detection systems use static rule definitions. Drawing data from either publicly available or paid databases, they rely on humans to define the criteria that will trigger an alert. A unique challenge posed by IDS systems is the sheer volume and diversity of data that must be processed and analyzed in a way so as not to disrupt communication and data traffic. In most cases, an IDS uses mirrored traffic instead of the actual real-time flow, although this leads to a slower response time.

The solution suggested in the paper is to use models similar to the CatBoost classification model that was additionally improved by an optimization algorithm. With this method, the machine learning model is able to learn what the typical activities of users and the network are, while any deviation from those standards would produce an alarm and force an investigation by an analyst. The procedure is highly effective in detecting the traitor intruder type. Since traitors are able to mask their malicious activity within their regular actions naturally, a holistic approach is necessary to face the challenges of spotting an obvious pattern in the activity log. Accordingly, if the traitor has sufficient technical expertise or privilege, it becomes trivial for him to bypass publicly available static rules.

LightGBM (light gradient boosting machine) can be used to deal with the large volume of data generated by the IDS. A greater data volume necessitates greater sampling, which in turn demands more resources. LightGBM communicates and compares the local best split points to find the single best one. Further, this reduces the sample sizes and gives better performance compared to similar algorithms. This comes with the compromise that configuring for such results is more challenging [4].

Extreme Gradient Boosting (XGBoost) is another potential technique that is applicable to the field of cybersecurity, specifically in the context of real-time systems, under which most security and industrial systems are classified. The key advantage is the ability to handle varied and complex datasets, which have to be contrasted and compared to come to a proper conclusion regarding the validity of a threat or alert [5].

3. METAHEURISTIC OPTIMIZATION SOLUTIONS

Metaheuristic optimization is a form of optimization that does not guarantee a globally optimal solution, although, in the field of cybersecurity, it could be argued that there is no such thing. As the field of optimization is still rapidly developing with many schools of thought based upon different approaches, this paper looks only into the ones who sought inspiration in animals and natural processes – algorithms like particle swarm optimization (PSO), artificial bee colony (ABC), firefly algorithm (FA), bat algorithm (BA), chimp optimization algorithm (ChOA), etc. There are other solutions that are inspired by mathematical approaches, such as the sine cosine algorithm, where the mathematical functions are used to bound values through each iteration and, at



the same time, to ensure smaller steps [6]. Swarm-based algorithms, on average, place a higher emphasis on exploration and, for that reason, will be considered above other solutions.

For example, ChOA was developed with the intent to emulate the hunting techniques and tactics employed by actual chimpanzees. This draws natural parallels with the red and blue team dynamics seen in cybersecurity when the blue team is hunting down the red team and employing various agents/tools in order to corner and identify the opponents. Combating threat actors requires multi-dimensional searches and trans-media analysis of log sources, which is also the strength of this algorithm. ChOA can be further enhanced with the incorporation of quasi-adaptive learning (QRL). In this way, the solution is split into two halves, where only one is created by applying the conventional unmodified ChOA, while the other uses QRL. Additionally, there is a rollback functionality added, meaning that in case of stagnation during an iteration, the altered ChOA can go back to a previous one to attempt a new solution. This novel algorithm was therefore named iteration stagnation aware ChOA (ISA-ChOA) [4]. An example of a pseudocode for ISA-ChOA is given in Listing 1.

Another potential swarm behavior algorithm that can be used is the firefly algorithm (FA). Like ChOA it requires a form of modification to produce a performance superior to the original one. In its unmodified form, FA simulates the behavior of a swarm of fireflies that are mutually attracted based on their luminosity.

This luminosity is a fitness function of their position. The fireflies with lower brightness intensity will be attracted to the ones of higher intensity; if there are no such, they will move around randomly. This cycle of attraction is repeated until achieving a satisfactory convergence or until enough time has passed (enough iterations have been reached).

One of the drawbacks of FA is a lack of exploration power. The initial distribution and positioning of the fireflies are randomly generated, and this kind of randomness often leads to immediate convergence towards sub-optimal solutions. To add more diversity and enhance the ability of the algorithm to seek optimal solutions, the following modifications were added:

- Genetic operators – uniform crossover and Gaussian mutation
- Quasi-reflection – based learning mechanism
- Dynamically adjusting the step size parameter.

Genetic operators allow for diversification, both solutions and fireflies. By combining existing and random solutions, they make an improvement in exploration that leads to new solutions, among which a more optimal one is to be found [7].

Quasi-reflection works oppositely and is comparable to the form of proof by contradiction. It seeks out the opposite solution, using the inverse values of the ones in the fitness function, which have been proven to provide the current most optimal solution.

```

Set maximum number of agents N
Set stagnation criteria st
Produce N/2 of the population P
Produce remaining agents by applying QRL
Separate agents in to simulated chimp colonies
while t<T do
    Evaluate agent fitness
    Use colony appropriate strategy to update c, f and m
    for Each search agent a in P do
        for each search agent s do
            Determine appropriate search strategy
            Update agent position
        end for
    end for
    Check for stagnation
    if Stagnation confirmed then
        Apply soft rollback
    else
        Store solutions for soft rollback
    end if
end while

```

Listing 1. ISA-ChOA optimizer pseudo code [4]



Dynamically adjusting the step size parameter in this context means reducing it, i.e., the algorithm converges on a solution, thereby altering the scope from a global to a local search. This value is bounded by a minimum that the parameter is not allowed to exceed.

These improvements create a modified version of the algorithm that is referred to as genetic operators quasi-reflected FA (GOQRFA Listing 2). It allows for the exploration of previously unknown or under-promising solutions in the hope of finding an actual promising and optimal one. The replacement for the worst-performing solution is created by mutating it or merging it with another random solution. Every subsequent iteration continuously elevates this lower boundary of the worst solution and exploits the narrowing spectrum of the potentially optimal one [8].

The FA is still not ideal even when modified since it has a potential worst-case complexity of $O(N) + O(N^2 \cdot T)$, where in the equation N is the number of fireflies and T is the number of iterations. To avoid this complexity, the initial data must be considered and analyzed in terms of distribution. For this to be effective, the sorting algorithm must have a baseline complexity that is lower than that of the worst-case scenario for FA [9].

While MFA/GOQRFA solves the issue of over-localized searches and under-exploration, the red fox algorithm (RFO) prioritizes a global search. Unlike FA, where the individual fireflies will group up together, foxes will seek out new “food sources”. In this context, the food source is a potential solution, the foxes will follow in a certain direction and evaluate it with a fitness function [10]. If the newly discovered value is better than

```
Metaheuristics parameter's values initial setting
Population P production
P evaluation with regard to the fitness function
for i = 1 to max iteration count do
  for every individual do
    for every better individual do
      if individual is better then
        Obtain attraction in terms of distance
        Adjust position toward the better individual
      end if
    end for
  Evaluate and update individuals in population P
  Produce novel solution by applying genetic crossover mechanism
  Subject novel solution to mutation
  Replace the worst-performing individual with a novel generated solution
end for
Return top-performing solution
end for
```

Listing 2. Modified firefly algorithm pseudocode [10]

```
Generate 1/2 of the optimization group P
Apply QRL to initialize the later portion of solutions.
Define  $\theta=0.8$ 
while  $T>t$  do
  Assess group fitness
  Select a random value for  $\psi$ 
  if  $\psi>\theta$  then
    Apply FA search
  else
    Apply RFO search
  end if
   $\theta=\theta-0.4$ 
end while
```

Listing 3. HARFO pseudocode [3]



the previous one, the fox will stay in place; if not, it will return to their prior position. This is comparable to the behavior of the chimps in ISA-ChOA.

On the other hand, RFOs can struggle to find a balance between exploitation and exploration. Too much exploitation makes it run into similar problems as FA, while too much exploration leads to diminishing returns, where each iteration leads to no significant improvements or new solutions. These issues can be addressed in the same way as in FA, i.e., with quasi-reflexive learning. One-half of the initial population is initialized in the same way as in unmodified RFO, while the other half is initialized through quasi-reflexive learning, through a random function that operates within the upper and lower bounds of the region being searched. This version of RFO is known as the hybrid adaptive RFO (HARFO), which is shown in the form of a pseudocode in Listing 3 [3].

This is a non-exhaustive list of swarm-based algorithms. An analysis of other optimization solutions like bat algorithm, artificial bee colony, lizard algorithm, etc., is left for subsequent papers. Just like algorithms, the search for a solution must be limited to a certain scope, and first, it is necessary to find a locally optimal solution before expanding and evaluating other possibilities.

4. ANALYSIS

In the analysis, the paper considers three points deriving from the previously listed algorithms and their implementations: the challenges of training models with adequate realistic data; the proven improvement of outcomes when employing the optimization algorithms; and the cost and requirements that would precede any implementation.

The first issue created using these optimization models is the acquisition of and training on real-world data [3]. Much of the cited analysis was done on simulated data and based on publicly available datasets. However, as the security environment is different and uniquely designed for a certain system or corporation, that means that the variety, quality, and details within logs and datasets will greatly differ on a case-by-case basis. An additional concern is whether the system handles information from third parties or from users who have not consented to have their data used to train a machine learning model. Furthermore, as blind application of a model will not yield high-quality results and the precision required in cybersecurity, the full understanding of a model's

decision-making must precede any evaluation of a model's performance. There is a clear separation between simpler models, which can be reduced to decision trees, and more complex models, which would require mathematical analysis and detailed interpretations [5].

Secondly, there is an empirically and experimentally verified improvement in performance when optimization algorithms are used on machine learning models. Specifically in the case of comparative analysis and multi-class classification of various types of attacks against IoT systems, "a custom-altered optimizer (..) resulting in the best performing models (..) attained a supreme accuracy level of 99.83%" [4]. These results show a large gap between non-optimized and optimized machine learning models. With the benefits of optimization clearly established, the cost and requirements have to be evaluated as well.

Thirdly, real-time mitigation as well as real-time attack detection is a key concern since the impact on customers, third parties, and alike must be minimal to nonexistent. This is a requirement for a modern and competitive security solution. The optimization algorithms would have to be evaluated in a real-world case, and that would mean analyzing real-world data, which is unpredictable and ever-changing based on trends and events outside of the programmer's or security analyst's control. Also, it would be in addition to the hardware requirements in terms of data throughput and processing requirements. During the learning stages of the algorithm, the data would have to be doubled to prevent any degradation or damage in case of false positives until the model establishes a baseline of activity.

The conclusion of the analysis emphasizes that the empirical data gained through simulation is twofold. First, it shows that the use of metaheuristic optimization algorithms guarantees better results, and second, that the practical and logistical concerns need to be resolved before any testing could be performed in a practical scenario or environment. Furthermore, either the legality of data harvesting would have to be legislated, or the data would need to be sufficiently anonymized. This paper does not consider anonymity to be absolute or even theoretically possible if there is enough data from a user available.



5. CONCLUSION

ISA-ChOA, GOQRFA, and HARFO are presented as potentially good solutions to the problems faced in the dynamic landscape of cybersecurity. By emulating a hunting scenario, solutions like ISA-ChOA and HARFO are appropriate for dealing with masqueraders since they lean more towards exploration and testing out new potential solutions. Meanwhile, GOQRFA leans more towards detecting both the accidental malicious actor and the traitor, by highlighting and focusing on the specific areas of interest. Thus, it can help identify exceptionally vulnerable areas or highly trafficked systems that can be leveraged during an exfiltration.

Since there is no single universal algorithm that can identify every single type of attacker efficiently, this paper proposes a hybrid solution defined as a parallel use of the three suggested metaheuristic optimization algorithms (FA, ChOA, RFO). The greatest risk to an IDS is a threat that it has no rules for – a zero-day exploit not yet in any database. By employing AI, a system is able to construct its database, which is both wholly unique to the company and has a use case relevant to the industry it operates. Clear optimization is a requirement for avoiding unnecessary waste of resources or inefficient searches that would interfere with regular operations. Although the investment in resources is non-trivial, the benefits and improved results are proven and verified. Therefore, algorithms like ISA-ChOA, GOQRFA, and HARFO present a potential solution for efficient intruder detection systems.

REFERENCES

- [1] L. Liu, O. De Vel, Q.-L. Han, J. Zhang and Y. Xiang, "Detecting and preventing cyber insider threats: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 1397-1417, 2018. doi: 10.1109/COMST.2018.2800740.
- [2] F. L. Greitzer, J. Strozer, S. Cohen, J. Bergey, J. Cowley, A. Moore and D. Mundie, "Unintentional insider threat: contributing factors, observables, and mitigation strategies," in *2014 47th Hawaii International Conference on System Sciences*, Waikoloa, HI, USA, 2014. doi: 10.1109/HICSS.2014.256.
- [3] D. Mladenovic, M. Antonijevic, L. Jovanovic, V. Simic, M. Zivkovic, N. Bacanin, T. Zivkovic and J. Perisic, "Sentiment classification for insider threat identification using metaheuristic optimized machine learning classifiers," *Scientific Reports*, vol. 14, no. 1, p. 25731, 2024. doi:10.1038/s41598-024-77240-w
- [4] M. Antonijevic, M. Zivkovic, M. Djuric Jovicic, B. Nikolic, J. Perisic, M. Milovanovic, L. Jovanovic, M. Abdel-Salam and N. Bacanin, "Intrusion detection in metaverse environment internet of things systems by metaheuristics tuned two level framework," *Scientific Reports*, vol. 15, no. 1, p. 3555, 2025. doi:10.1038/s41598-025-88135-9
- [5] P. Dakic, M. Zivkovic, L. Jovanovic, N. Bacanin, M. Antonijevic, J. Kaljevic and V. Simic, "Intrusion detection using metaheuristic optimization within IoT/IIoT systems and software of autonomous vehicles," *Scientific Reports*, vol. 14, no. 1, p. 22884, 2024. doi:10.1038/s41598-024-73932-5
- [6] M. Dobrojevic, M. Zivkovic, A. Chhabra, N. S. Sani, N. Bacanin and M. M. Amin, "Addressing internet of things security by enhanced sine cosine metaheuristics tuned hybrid machine learning model and results interpretation based on shap approach," *PeerJ Computer Science*, vol. 9, p. e1405, 2023. doi: 10.7717/peerj-cs.1405
- [7] S. Mirjalili and S. Mirjalili, "Genetic algorithm," *Evolutionary algorithms and neural networks: Theory and applications*, pp. 43-55, 2019. source: <https://www.iieta.org/journals/jesa/paper/10.18280/jesa.560601>
- [8] N. Bacanin, M. Zivkovic, T. Bezdan, K. Venkatachalam and M. Abouhawwash, "Modified firefly algorithm for workflow scheduling in cloud-edge environment," *Neural computing and applications*, vol. 34, no. 11, pp. 9043-9068, 2022. DOI:10.1007/s00521-022-06925-y
- [9] M. Zivkovic, M. Tair, N. Bacanin, S. Hubalovsky and P. Trojovsky, "Novel hybrid firefly algorithm: An application to enhance XGBoost tuning for intrusion detection classification," *PeerJ Computer Science*, vol. 8, p. e956, 2022. DOI:10.7717/peerj-cs.956
- [10] N. Savanovic, A. Toskovic, A. Petrovic, M. Zivkovic, R. Damasevicius, L. Jovanovic, N. Bacanin and B. Nikolic, "Intrusion detection in healthcare 4.0 internet of things systems via metaheuristics optimized machine learning," *Sustainability*, vol. 15, no. 16, p. 12563, 2023. doi:10.1038/s41598-024-73932-5



FEDERATED LEARNING SETTING FOR E-LEARNING COURSE RECOMMENDATIONS

Miloš Jolović*,
[0009-0003-6580-2039]

Dušan Kostić,
[0009-0000-3432-6074]

Aleksandar Joksimović,
[0009-0008-5711-7636]

Talib Tahirović,
[0009-0004-9515-8237]

Petar Lukovac
[0000-0003-4561-8886]

University of Belgrade,
Faculty of Organizational Sciences,
Belgrade, Serbia

Abstract:

The main research problems addressed in this article refer to the complexity of maintaining Learning Management Systems, ensuring data privacy throughout any analysis of that data, and personalizing learning, which can be a task requiring significant resources. The research aims to provide an answer that can address these problems through a Federated Learning setting, enabling cross-institutional cooperation and retaining the data in its place of origin. The research includes a simulation of such a Federated learning setting, which proved to be very interesting for identifying future challenges and directions for a tangible, real-world application. The simulation was built with a dataset comprised of students' grades and interests in a first-year mandatory subject, E-business, taught at the University of Belgrade, Faculty of Organizational Sciences. This dataset was suitable for building a recommender system that can produce an intelligent suggestion for an elective course for each student individually based on their interests and academic achievements.

Keywords:

Federated Learning, E-learning, Personalized Learning.

INTRODUCTION

Recent advances in the generation of immense volumes of data have raised concerns about leveraging the generated data in a privacy-preserving manner. Consequently, a Federated Learning (hereinafter: FL) setting has emerged. The FL setting is a distributed machine learning technique where models are brought to local data on each node of the network, instead of centralizing vast volumes of data [1]. Several authors have proposed the use of FL in education for data analysis [2] [3] or for detecting dropout rates [4]. One of the potential applications of the FL setting is training models across academic institutions to develop and sustain personalized learning. Personalized learning can be defined as a complex activity that considers individual needs and goals in the process of learning [5]. This research paper aims to develop a FL simulation for making intelligent recommendations for students regarding choosing an appropriate elective course depending on their previous knowledge and interests.

Correspondence:

Miloš Jolović

e-mail:

milos.jolovic@elab.rs





2. RELATED WORK

The term FL was first introduced in [6] as a way to distribute the process of training a machine learning algorithm by a federation of clients, such as mobile devices or several different companies and institutions. The term FL has since evolved into a machine learning setting in which client nodes collaborate in training a machine learning model on their local data, with a central server orchestrating the training process [7].

The training process of a machine learning algorithm via FL can be described in four steps. Firstly, the server initializes a global machine learning model. Next, client nodes download the global model and train it on their local data. Then, client nodes send back the model parameters to the server node. Finally, the server aggregates the parameters using the FL algorithm and updates the global model. These steps are repeated until the model converges [8].

The FL paradigm has three different categories, depending on how data is partitioned in both feature and sample space: Vertical FL, Horizontal FL, and Federated Transfer Learning. Horizontal FL refers to a situation where clients hold data with the same features and different samples. Vertical FL refers to a situation where clients hold data with different features and share samples. Federated Transfer Learning refers to a situation where datasets differ in feature and sample spaces with limited overlaps [9] [10].

FL can also be categorized based on the amount of decentralization between the nodes as centralized (CFL), decentralized (DFL) and semi-decentralized FL (SDFL). In DFL, participants perform all four steps of the FL process independently, SDFL participants perform the first

three steps independently, the aggregator node handles the aggregation process and then passes the aggregator role to a new node in the next iteration of training. Centralized FL functions between a server node and several client nodes, as described in Figure 1 [11].

The justification for usage of the FL setting can be summarized as follows. First, machine learning algorithms can be trained on separate data silos (e.g., several medical institutions) in a privacy-preserving manner, without the need for any silo to share its local data. Second, some data sources provide a large amount of real-time data, making it more efficient to move the model to the data rather than vice-versa. Third, many legal regulations can make sensitive data hard to move from the place of origin [12] [13] [14] [15].

From a practical perspective FL setting can be used in healthcare informatics [16] [17], and the banking industry in several directions, such as assessing credit risks [18], open banking [19] or credit fraud detection [20]. Practical applications can be found in IoT systems [21] [22], wireless communications [23], the automotive industry [24] Etc.

More specifically, practical usage of FL can be found in an e-learning setting. The concept of e-learning can be described as an educational process that leverages digital platforms and resources to facilitate the learning process and educational resources [25]. Research conducted by [26] states that students often use the same edge devices (e.g., PCs, laptops, mobile phones) for both entertainment and study assignments, which can cause problems with maintaining focus on study tasks. The authors propose an FL architecture that collects data from students' edge devices in a privacy-preserving way to train a classification model for students' on-screen

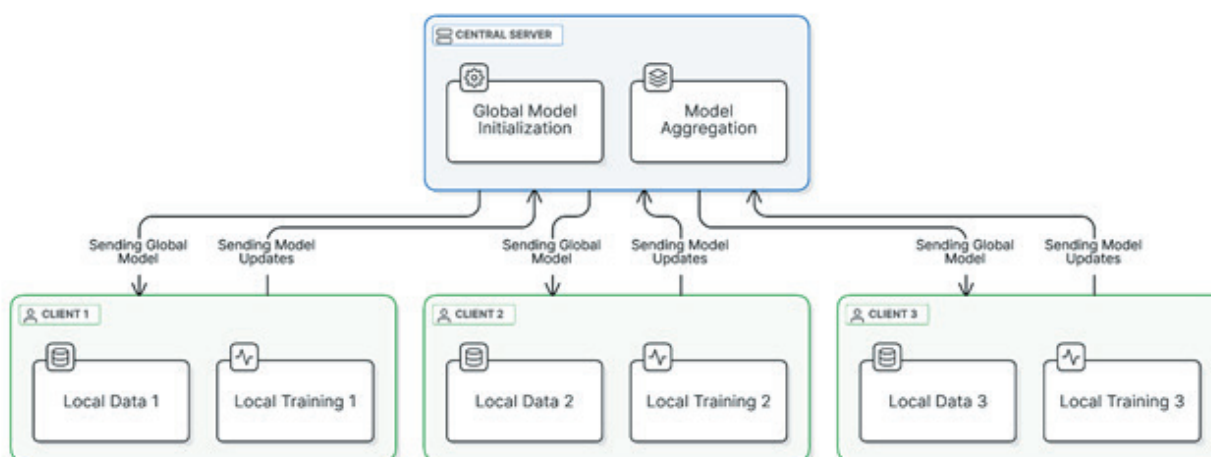


Figure 1. Federated learning process



time that detects situations in which study tasks lose students' attention due to entertainment. Research [27] recognizes the significance of personalized learning and the possibilities of implementing it in an e-learning setting, where security and data privacy may arise. Recommendation systems trained via an FL architecture are proposed as a solution for implementing personalized learning in a secure and privacy-preserving way.

The simulation of an FL setting for this paper was implemented through the Flower Framework. Flower is an FL framework that provides a unified approach to FL, analytics, and evaluation that can be applied to federating any workload or machine learning framework [28]. This research paper proposes a simulation of an FL setting in e-learning, specifically building a recommendation model to suggest elective courses to students depending on their previous interests and academic accomplishments. The simulation environment represents an opportunity to identify the benefits of FL setting in e-learning, such as cross-institutional cooperation between academic institutions without the need to share raw data, learning personalization and possible integration with Learning Management Systems.

3. METHODOLOGY

Data used for this research paper was collected from students enrolled in the Information Systems and Technologies undergraduate study program at the Faculty of Organizational Sciences, University of Belgrade. This study program focuses on applying computer science to construct solutions for business-oriented problems. More precisely, the dataset comprised students' grades and topics of project-based activities from the first-year E-business course. E-business provides students with both theoretical and practical, introductory knowledge of building web-based applications. Course assessment consists of mandatory closed-format tests once a week, four project-based homework assignments and one exploratory essay designed to encourage students to explore state-of-the-art technologies regarding various topics of computer science such as blockchain, artificial intelligence, big data, cloud technologies, IoT solutions, etc. The procedure is described below.

Research began with collecting data from Moodle Learning Management System (LMS). The Department of E-business utilizes Moodle throughout the teaching process for posting lecture resources, communication with students, assessments and grading. As mentioned before, the dataset consists of students' grades on vari-

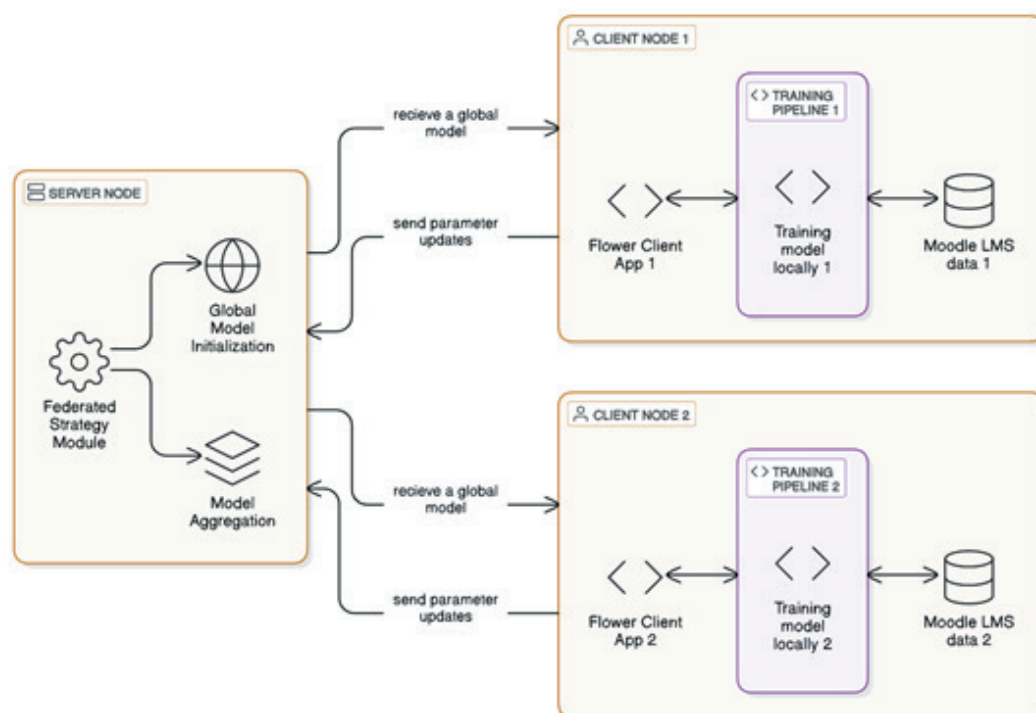


Figure 2. Federated learning project architecture



ous homework assignments and tests, with topics of their exploratory essays. The collected data was used to build an intelligent recommendation system that suggests elective courses for students based on their grades and previous interests expressed in essays done on the E-business course. The Dataset contains 609 student records.

In the process of building a recommender model data preparation involved replacing missing values with zeroes – assigning zero points to students who did not complete a particular assignment. Students who did not successfully complete the course were removed from the dataset. Numerical columns were normalized, and students were grouped into three clusters based on their performance in the E-business course. These clusters served as recommendations for the level of elective courses that students should take. The clustering was performed using the K-Means algorithm, implemented with the scikit-learn library. Data containing elective courses topics and descriptions was processed using the Stanza library and transformed into TF-IDF matrices. The same procedure was applied to students' essay topics and their respective categories. Cosine similarity was computed between the matrices representing course descriptions and students' essays. The three highest similarity scores were used to generate personalized elective course recommendations for each student.

The architecture of FL was built with Flower Framework using two client nodes. Each client node received a global model initialized on the server node and returned updates to the global model. The server node aggregated the model updates using the Federated Averaging strategy.

4. RESULTS

As a result of training on students' data, three clusters emerged:

- Cluster 0 – students with good performance and a minor lack of theoretical knowledge.
- Cluster 1 – students that lack both theoretical and practical knowledge and should cover the basics again.
- Cluster 2 – students that excel in the E-business course and have a great understanding of both theoretical and practical knowledge.

Students from Cluster 0 were recommended intermediate-level elective courses, Cluster 1 students were advised to choose elective courses which will cover the basics again, and students from Cluster 2 were encouraged to enroll in advanced elective courses.

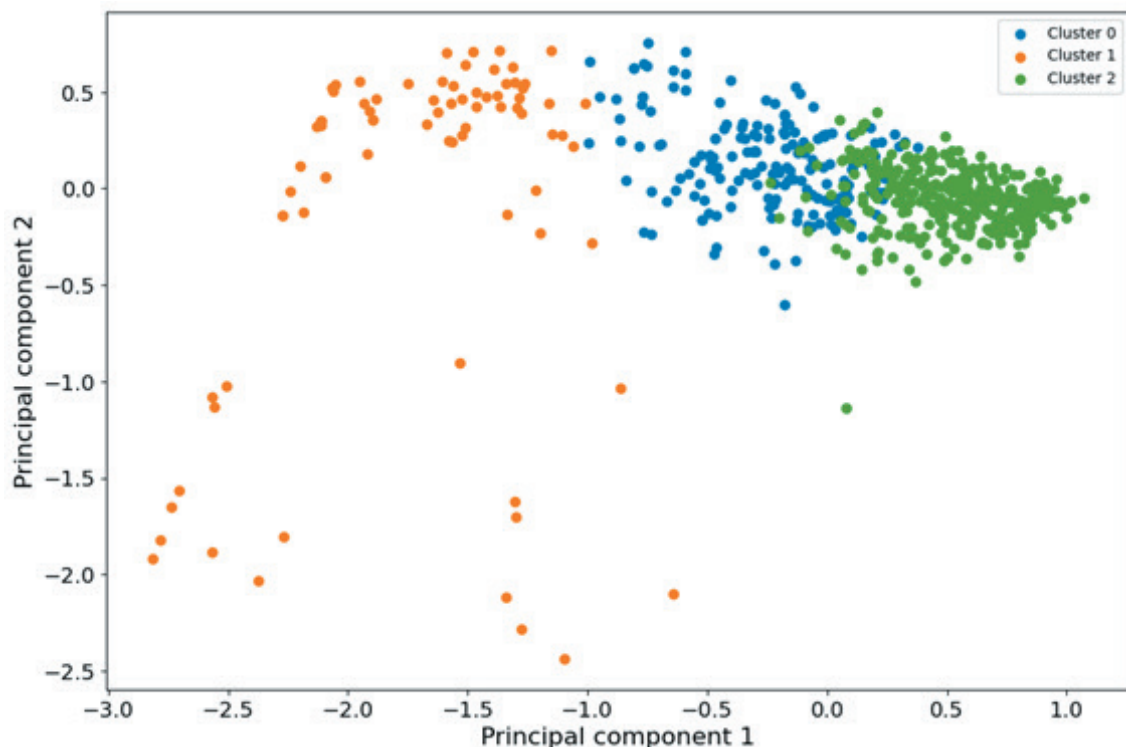


Figure 3. Student clusters emerged from the K-Means clustering algorithm



The top three elective course recommendations were generated for each student based on the similarity between their exploratory essay topics and elective course descriptions. For instance, students who wrote about IoT in their essays were recommended IoT-based elective courses.

5. CONCLUSION

This pilot research demonstrated a simulation of an FL setting implemented with Flower Framework. The study aimed to highlight the potential of applying federated learning in possible cross-institutional collaboration and personalization of e-learning. The research also proves to be valuable in identifying potential challenges that may arise during the development of a true-to-life federated architecture.

Future research could focus on expanding cross-institutional cooperation by building and deploying a fully functional FL system across multiple universities or faculties. This system could integrate real-time student progress tracking and incorporate more complex metrics. However, several challenges must be addressed, including security risks in federated architectures, data heterogeneity across institutions (client nodes), and the need for a custom aggregation strategy tailored to varying data quality and volume at each institution.

REFERENCES

- [1] K. M. Jawadur Rahman et al., "Challenges, Applications and Design Aspects of Federated Learning: A Survey," *IEEE Access*, vol. 9, pp. 124682–124700, 2021, doi: 10.1109/ACCESS.2021.3111118.
- [2] C. Fachola, A. Tornarí, P. Bermolen, G. Capdehourat, L. Etcheverry, and M. I. Fariello, "Federated Learning for Data Analytics in Education," *Data 2023*, Vol. 8, Page 43, vol. 8, no. 2, p. 43, Feb. 2023, doi: 10.3390/DATA8020043.
- [3] S. Guo, D. Zeng, and S. Dong, "Pedagogical data analysis via federated learning toward education 4.0," *American journal of education and information technology*, vol. 4, no. 2, pp. 56–65, 2020, doi: 10.11648/J.AJEIT.20200402.13.
- [4] T. Zhang et al., "Enhancing Dropout Prediction in Distributed Educational Data Using Learning Pattern Awareness: A Federated Learning Approach," *Mathematics 2023*, Vol. 11, Page 4977, vol. 11, no. 24, p. 4977, Dec. 2023, doi: 10.3390/MATH11244977.
- [5] A. Shemshack and J. M. Spector, "A systematic literature review of personalized learning terms," *Smart Learning Environments*, vol. 7, no. 1, pp. 1–20, Dec. 2020, doi: 10.1186/S40561-020-00140-9/TABLES/4.
- [6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Apr. 10, 2017, *PMLR*. Accessed: Feb. 24, 2025. [Online]. Available: <https://proceedings.mlr.press/v54/mcmahan17a.html>
- [7] P. Kairouz et al., "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, Jun. 2021, doi: 10.1561/22000000083.
- [8] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, and W. Zhang, "A survey on federated learning: challenges and applications," *International Journal of Machine Learning and Cybernetics 2022 14:2*, vol. 14, no. 2, pp. 513–535, Nov. 2022, doi: 10.1007/S13042-022-01647-Y.
- [9] Y. Liu et al., "Vertical Federated Learning: Concepts, Advances, and Challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3615–3634, Jul. 2024, doi: 10.1109/TKDE.2024.3352628.
- [10] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards Personalized Federated Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9587–9603, Dec. 2023, doi: 10.1109/TNNLS.2022.3160699.
- [11] E. T. Martinez Beltran et al., "Decentralized Federated Learning: Fundamentals, State of the Art, Frameworks, Trends, and Challenges," *IEEE Communications Surveys and Tutorials*, vol. 25, no. 4, pp. 2983–3013, 2023, doi: 10.1109/COMST.2023.3315746.
- [12] Q. Li et al., "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3347–3366, Apr. 2023, doi: 10.1109/TKDE.2021.3124599.
- [13] T. R. Gadekallu, Q.-V. Pham, T. Huynh-The, S. Bhattacharya, P. K. R. Maddikunta, and M. Liyanage, "Federated Learning for Big Data: A Survey on Opportunities, Applications, and Future Directions," Oct. 2021, Accessed: Feb. 26, 2025. [Online]. Available: <https://arxiv.org/abs/2110.04160v2>
- [14] P. M. Mammen, "Federated Learning: Opportunities and Challenges," Jan. 2021, doi: 10.1145/nnnnnnnn.nnnnnnnn.
- [15] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. K. Guo, "Privacy preservation in federated learning: An insightful survey from the GDPR perspective," *Computers & Security*, vol. 110, p. 102402, Nov. 2021, doi: 10.1016/J.COSE.2021.102402.



- [16] N. Rieke et al., “The future of digital health with federated learning,” *npj Digital Medicine* 2020 3:1, vol. 3, no. 1, pp. 1–7, Sep. 2020, doi: 10.1038/s41746-020-00323-1.
- [17] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, “Federated Learning for Healthcare Informatics,” *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, Mar. 2021, doi: 10.1007/S41666-020-00082-4/TABLES/2.
- [18] D. Kawa, S. Punyani, P. Nayak, A. Karkera, and V. Jyotinagar, “Credit Risk Assessment from Combined Bank Records using Federated Learning,” *International Research Journal of Engineering and Technology*, p. 1355, 2008, Accessed: Feb. 26, 2025. [Online]. Available: <https://www.rbi.org.in/Scripts/NotificationUser.aspx>
- [19] G. Long, Y. Tan, J. Jiang, and C. Zhang, “Federated Learning for Open Banking,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12500 LNCS, pp. 240–254, 2020, doi: 10.1007/978-3-030-63076-8_17.
- [20] W. Yang, Y. Zhang, K. Ye, L. Li, and C. Z. Xu, “FFD: A federated learning based method for credit card fraud detection,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11514 LNCS, pp. 18–32, 2019, doi: 10.1007/978-3-030-23551-2_2/TABLES/4.
- [21] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, “Federated Learning for Internet of Things: A Comprehensive Survey,” *IEEE Communications Surveys and Tutorials*, vol. 23, no. 3, pp. 1622–1658, Jul. 2021, doi: 10.1109/COMST.2021.3075439.
- [22] V. Rey, P. M. Sánchez Sánchez, A. Huertas Celdrán, and G. Bovet, “Federated learning for malware detection in IoT devices,” *Computer Networks*, vol. 204, p. 108693, Feb. 2022, doi: 10.1016/J.COMNET.2021.108693.
- [23] S. Niknam, H. S. Dhillon, and J. H. Reed, “Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges,” *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, Jun. 2020, doi: 10.1109/MCOM.001.1900461.
- [24] H. Zhang, J. Bosch, and H. H. Olsson, “Real-time end-to-end federated learning: An automotive case study,” *Proceedings - 2021 IEEE 45th Annual Computers, Software, and Applications Conference, COMPSAC 2021*, pp. 459–468, Jul. 2021, doi: 10.1109/COMPSAC51774.2021.00070/VIDEO.
- [25] S. Bashir and A. L. Lapshun, “E-learning future trends in higher education in the 2020s and beyond,” *Cogent Education*, vol. 12, no. 1, Dec. 2025, doi: 10.1080/2331186X.2024.2445331.
- [26] D. Mistry, M. F. Mridha, M. Safran, S. Alfarhood, A. K. Saha, and D. Che, “Privacy-Preserving On-Screen Activity Tracking and Classification in E-Learning Using Federated Learning,” *IEEE Access*, vol. 11, pp. 79315–79329, 2023, doi: 10.1109/ACCESS.2023.3299331.
- [27] D. Javeed, M. S. Saeed, P. Kumar, A. Jolfaei, S. Islam, and A. K. M. N. Islam, “Federated Learning-Based Personalized Recommendation Systems: An Overview on Security and Privacy Challenges,” *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2618–2627, Feb. 2024, doi: 10.1109/TCE.2023.3318754.
- [28] D. J. Beutel et al., “Flower: A Friendly Federated Learning Research Framework,” Jul. 2020, Accessed: Feb. 20, 2025. [Online]. Available: <https://arxiv.org/abs/2007.14390v5>



EFFECTS OF ADAM OPTIMIZER VARIANTS ON BRAIN TUMOR SEGMENTATION TASK

Samson Offorjindu,
[0009-0001-2928-0654]

Marina Marjanović,
[0000-0002-9928-6269]

Timea Bezdan*
[0000-0001-6938-6974]

Singidunum University,
Belgrade, Srebia

Abstract:

In medical image analysis, accurately segmenting brain tumors is still very challenging, motivating researchers to explore advanced deep-learning methods. While U-Net models have produced promising results, improving their performance through optimized training techniques is still necessary. Given that Adam is commonly used as the default optimizer in such tasks, our study explores the impact of different Adam optimizer variants on U-Net performance using the well-known BraTS 2020 dataset. We evaluated Adam, AdamW, Adagrad, Adamax, Adafactor, and RMSprop optimizers, comparing their performance using key metrics such as training loss, validation loss, F-score, Intersection over Union (IoU), precision, and recall. The obtained results show that Adamax achieves the highest F-score (0.8120) and IoU score, demonstrating superior performance in segmenting tumor regions in medical images; AdamW also showed strong results with lower training and validation losses, as well as good precision and recall, highlighting its efficiency and accuracy. These findings emphasize the importance of selecting the right optimizer for Li-Net-based brain tumor segmentation and encourage further exploration into optimized training strategies in medical image analysis.

Keywords:

Artificial Intelligence, Computer Vision, Medical Image Segmentation, Convolutional Neural Network, Deep Learning.

INTRODUCTION

Brain Tumor Segmentation [1-5] has always focused on evaluating state-of-the-art methods for segmenting brain tumors in multimodal magnetic resonance imaging scans (MRIs). The BraTS 2020 dataset, a widely used benchmark in the field, utilizes multi-dimensional pre-operative MRI scans and primarily focuses on the segmentation task of intrinsically heterogeneous brain tumors, likely in appearance, shape, and histology, namely gliomas. Additionally, the dataset includes clinical information such as overall survival, the clinical assessment of disease progression, and uncertainty estimation for expected tumor subregions; BraTS multimodal imaging data include native T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes, provided in Neuroimaging

Correspondence:

Timea Bezdan

e-mail:

tbezdan@singidunum.ac.rs





Informatics Technology Initiative - NIfTI (.nii) format, these scans have been collected from various clinical procedures and imaging scanners at multiple institutions, all contributing to the dataset.

This research paper utilized a pre-trained deep-learning model for the brain tumor segmentation task using MRI scans. The model is trained on the BraTS 2020 dataset and is designed to identify and segment the different tumor subregions (whole tumor, tumor core, and enhancing tumor). The primary goal of the experiment is to evaluate the effectiveness of the other Adam optimizer variants in performing accurate segmentations on MRI scans while maintaining recall and precision without additional training. In contrast, we seek to determine whether and how the variations in the Adam optimizer's parameters affect the model's ability to accurately segment tumor subregions while maintaining a reasonable true-positive and true-negative rate. This study's central research question is of paramount importance: To what extent do different variants of the Adam optimizer influence the segmentation performance of a U-Net model (Figure 1) on the BraTS 2020 brain tumor dataset? This question is not only significant for our research but also for the broader field of medical image analysis and deep learning.

The Adam optimizer [6] is a widely recognized algorithm in deep learning and often the default choice; the method leverages the advantages of AdaGrad [7] and RMSProp [8], adapting the learning rates for each parameter based on their historical gradients. However, subtle variations in its parameters influence its convergence behavior and generalization performance. These differences often lead to variations in segmentation accuracy, mostly in complex medical image segmentation tasks where precise boundary outline is critical. We aim to provide a comparative analysis of six popular Adam optimizer variants: Adafactor [9], Adagrad, Adam, Adamax [10], AdamW, and RMSProp, specifically for our segmentation task. By examining their core mechanisms, strengths, and weaknesses, our analysis will provide insights into the suitability of each optimizer. More importantly, by leveraging a pre-trained model, we also aim to minimize computational costs while maintaining good segmentation accuracy, minimizing false rates, and retaining reasonable true positive rates. This research has practical implications for assisting radiologists in clinical decision-making, providing a practical and useable tool for the field of medical image analysis and deep learning.

The rest of this paper is organized as follows: Section 2 describes the methodology and model architecture. Section 3 presents the experimental setup, results, and discussion, and Section 4 concludes the paper with key findings and directions for future work.

2. METHODOLOGY

In digital image processing and computer vision, the process of partitioning an image into various multiple segments is known as the image segmentation task, where the goal is to simplify and transform the image representation to facilitate more efficient and accurate analysis. Image segmentation primarily identifies boundaries and objects, such as lines, curves, and other structures, within images. Image segmentation involves assigning a label to each pixel in an image so that pixels with the same label share a common set of characteristics. Our goal is to assign a unique label to each pixel, thereby outlining tumor boundaries in medical images.

The U-Net architecture is the most suitable architecture for our segmentation task due to its proven efficiency in medical image segmentation and its ability to handle limited training data, which results from several key characteristics:

- i. Its encoder-decoder structure, based on the fully convolutional neural network principles proposed in [11], effectively captures both high-level contextual information and low-level spatial details. The encoder extracts hierarchical features while the decoder reconstructs the segmentation map. Hence, a compelling feature hierarchy capture.
- ii. Implementing skip connections that bridge the encoder and decoder enables the propagation of the finely grained spatial information to higher-resolution layers. This is particularly vital in medical image segmentation, where precisely defining tumor boundaries is challenging. It enhances its robustness with limited data.
- iii. The output of pixel-wise segmentation maps directly addresses the need for precise, detailed tumor subregion delineation.

The core concept involves augmenting a conventional contracting network by substituting pooling operations with upsampling operators. Consequently, these layers enhance the output resolution. Moreover, this data enables a subsequent convolutional layer to construct a precise output. A key innovation of U-Net is the increased



density of feature channels in the upsampling pathway, facilitating the propagation of contextual information to higher-resolution layers. Therefore, the expansive pathway generates a U-shaped architecture and exhibits approximate symmetry with the contracting component. This tiling strategy is essential for applying the network to large images, as GPU memory limitations would otherwise constrain the resolution. The network utilizes only the valid portion of each convolution, excluding fully connected layers. The missing contextual information is extrapolated through input image mirroring to predict pixels in the image's border region.

Our research utilizes the proposed U-Net model for the tumor segmentation task. The model comprises three primary components: the ResNet encoder, a U-Net decoder, and a segmentation head. We utilized a ResNet encoder to extract a robust hierarchical feature. Its residual connections effectively address the vanishing gradient problem, enabling us to train deeper networks and capture more complex feature representations. The encoder begins with an initial convolutional layer, followed by batch normalization, ReLU activation, and max pooling. Subsequently, four sequential layers, composed of multiple Bottleneck blocks, perform feature extraction. Bottleneck blocks enhance computational efficiency while maintaining representational power. ResNet was used due to its strong ability to learn deep feature representations, which is particularly beneficial for analyzing complex medical images.

Moving on from the model, we will look at the Adam variant optimizers. Firstly, Adafactor offers significant memory efficiency, a crucial advantage when training large 2D U-Net models on high resolution images, potentially allowing for more complex architectures with limited resources. It provides adaptive learning rates that effectively handle the dataset's diverse intensity distributions and tumor characteristics. However, it possesses a slower convergence, which increases overall training times. Additionally, Adagrad adapts the learning rate for each parameter based on the cumulative history of gradients, which is beneficial for sparse features that might emerge. However, its weakness is the probability that the learning rate decays aggressively over time, leading to a slow convergence. Adam itself is a widely adopted optimizer that combines adaptive learning rates with momentum, making it computationally efficient; it possesses a deep, faster convergence due to its adjustments of individual parameter learning rates. Another variant of this optimizer uses the infinity norm and exhibits a stable behavior across different learning rates and gradient scales; Adamax exhibits robustness to extreme gradients, translating to a more stable training process. However, Adamax can still be sensitive to hyperparameter selection.

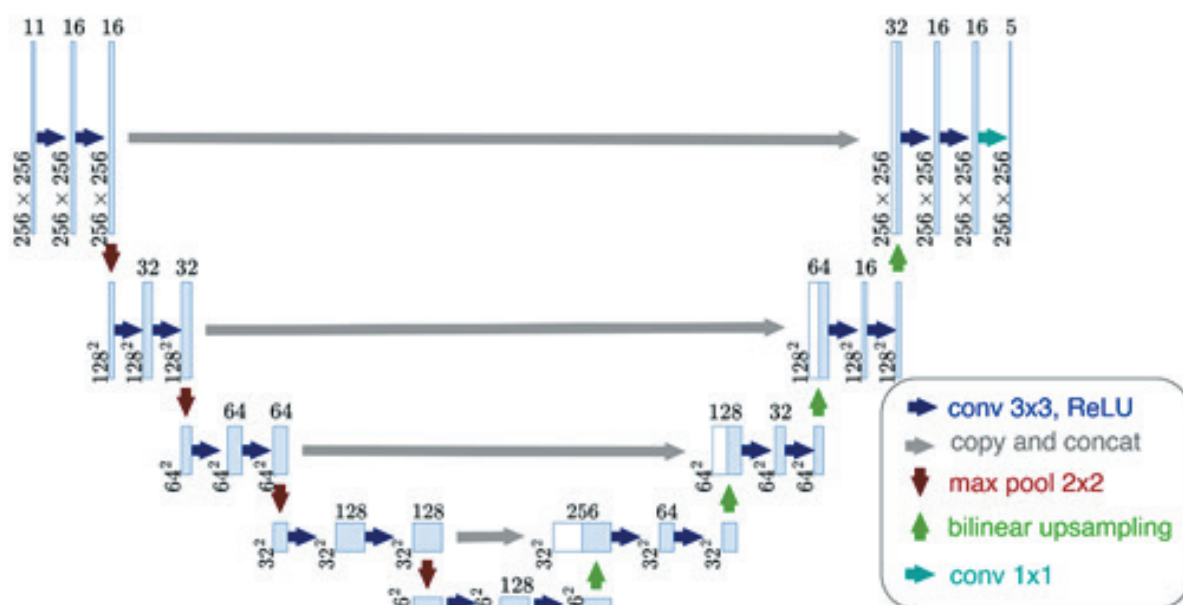


Figure 1. Model architecture



Regarding the Adam optimizer, AdamW improves upon the standard Adam optimizers by decoupling the weight decay process, leading to a more effective L2 regularization [12], giving it a higher performance potential for this task. Lastly, RMSprop is another optimizer that uses a moving average of squared gradients to normalize each parameter's learning rate, allowing it to handle parameters with diverse features. However, its performance is sensitive to the choice of hyperparameters and its slow convergence in most cases.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental setup in this study is designed to evaluate the performance of various Adam optimizer variants when training a U-Net model for brain tumor segmentation. The goal is to conduct a systematic comparison of these optimizers to determine how each influences segmentation accuracy, training dynamics, and generalization ability. By analyzing key performance metrics, we aim to reveal the strengths and limitations of each optimizer variant within the context of medical image segmentation.

3.1. EXPERIMENTAL SETUP AND DATASET

Dataset preparation and setup for training involved key stages to ensure proper data organization, pre-processing, and transformations. The dataset was initially sourced from the BraTS2020 training data, which is structured by extracting subject IDs from the CSV holding name-mapping data. Afterward, file paths are organized into separate categories corresponding to different MRI modalities, including the FLAIR, T1, T1CE, T2, and Segmentation masks (Figure 2), after which the currently structured dataset is subsequently stored in a CSV file for efficient access. The next stage involves loading the images using the NiBabel library that allows handling medical image files such as the one in question,

NIfTI (.nii). The images are passed onto a center cropping function, which ensures that the images maintain uniform dimensions across the dataset. Normalization is then applied to standardize pixel intensities, which contributes to stabilizing the training process, speeds up the training process, and improves the model's convergence.

Furthermore, data augmentation techniques were carried out to enhance model generalization. We utilized augmentation libraries from Albumentations and TorchVision for transformation operations such as image flipping, rotation, and contrast adjustments. These steps were taken to collectively ensure that the dataset is well-prepared for training our deep learning model - as represented in Table 1.

Additionally, the dataset was divided into training (60%), validation (20%), and test (20%) sets. Individual MRI slices were extracted and stored in .npy format for good loading and processing. The U-Net model was trained using the previously mentioned optimizers, each with a learning rate 0.0001. At the same time, dice loss was applied to the model's softmax2d output to optimize multi-class segmentation, and the batch size was set to 16 for training. We evaluated model performance using several metrics, including Intersection over Union (IoU) with a threshold of 0.5, F-score, precision, recall, training loss, and validation loss.

3.2. EXPERIMENTAL RESULTS AND DISCUSSION

This subsection describes the results of the experiments conducted using the Adam optimizer variants mentioned to train the U-Net model on the BraTS2020 dataset. The performance of these optimizer variants in terms of IoU, precision, recall, and F-score, as well as train and validation losses, provided insights into why Adamax had the best results. The comparative analysis is presented in Table 2.

Table 1. Parameter configuration

Parameter	Value
Batch size	16
Learning rate	1e-4
Loss function	DiceLoss
Activation function	ReLU
Epochs	200

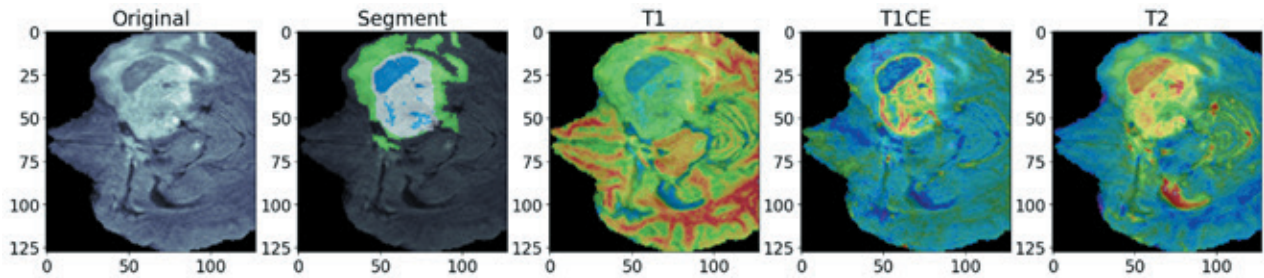


Figure 2. Sample modality visualizations

Table 2. Comparative analysis

Metric	Adam	AdamW	Adagrad	Adamax	Adafactor	RMSprop
Training loss	0.0025	0.0027	0.0171	0.0030	0.0097	0.0026
Validation loss	0.0108	0.0108	0.0183	0.0104	0.0124	0.0108
F-score	0.7964	0.7979	0.7105	0.8120	0.7661	0.8027
IoU-score	0.7087	0.7105	0.6125	0.7238	0.6743	0.7136
Precision	0.8711	0.8684	0.8670	0.8678	0.8713	0.8521
Recall	0.7810	0.7860	0.6869	0.8005	0.7486	0.7978

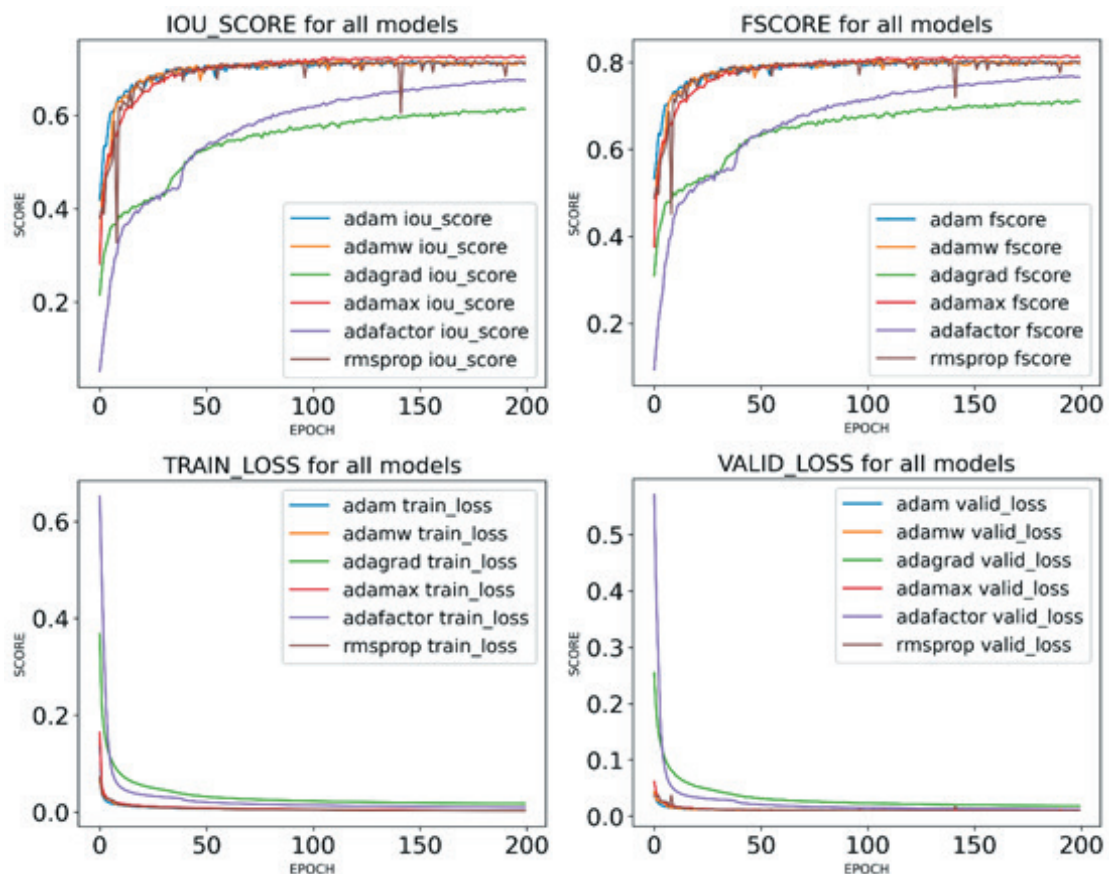


Figure 3-6. The first plot (top left) shows the Intersection of Union score across the epochs. The second plot (top right) displays the F-score with the third (on the right) and the validation loss across epochs



The experimental results (Figure 3-6) first showed a rapid initial loss reduction in both training and validation. The training dynamics across these optimizers showed a consistent pattern. Most of the methods show a rapid loss reduction, with convergence to stable loss values between 25 and 30 epochs; the training and validation curves mirrored each other with exceptional precision, including minimal overfitting and robust learning mechanisms, which suggests that the Adam optimizer family that is characterized by adaptive learning rates and momentum-based updates are very-well suited for the task. Within the first 75-100 epochs, we also observed stabilization for the IoU scores and precision and recall, after which the performance gains became increasingly marginal. This observation underscored the importance of early training stages and suggested that extended training may yield diminishing returns. However, Adamax's infinity norm variant appeared adept at handling parameter magnitudes variations, which should explain its performance advantage.

The precision scores, which represent the ability of the model to avoid false positives, were relatively similar across all three optimizers, with Adam achieving the highest precision (0.8711), followed closely by AdamW (0.8684) and Adamax (0.8678). This suggests that all three optimizers are reasonably effective at minimizing the prediction of tumor regions where none exist. However, Adamax demonstrated the highest recall (0.8005), followed by AdamW (0.7860) and Adam (0.7810). Recall measures the ability of the model to identify all actual positive cases, meaning Adamax was more successful in detecting all the tumor regions present in the BRATS2020 dataset compared to the other two optimizers. Overall, the initial analysis of the performance metrics suggests that Adamax outperformed both Adam and AdamW on key segmentation metrics and gener-

alization ability despite a slightly higher training loss. AdamW showed a marginal improvement over Adam in terms of the F-score and IoU score while maintaining the same validation loss.

While these results, as well as the predicted segmentations (Figure 7) seem rather compelling, they were not without limitations. Our analysis and experiments were conducted on a single dataset and experimental setup, which implies that generalizability will require further investigation and study. Nevertheless, the findings align closely with the existing context of adaptive optimization techniques, reinforcing the effectiveness of the Adam optimizer family. From a practical perspective, the results provide a clear recommendation. Adamax emerges as the preferred choice for tasks requiring a balanced performance, while the standard Adam serves as a robust alternative. The minimal variability between the tested optimizers suggests that other researchers and practitioners can confidently select from this optimizer family with a relatively low risk of significant performance degradation.

4. CONCLUSION

The findings of this study provide practical insights for researchers in medical image segmentation, particularly those using the BraTS dataset. Based on the obtained results, Adamax is the most effective optimizer among the other evaluated optimizers, offering superior segmentation accuracy and generalization capabilities. The outcome also highlights the potential benefits of Adamax's robustness to extreme gradients and noisy data, which are a widespread challenge in medical image analysis and many other deep-learning tasks.

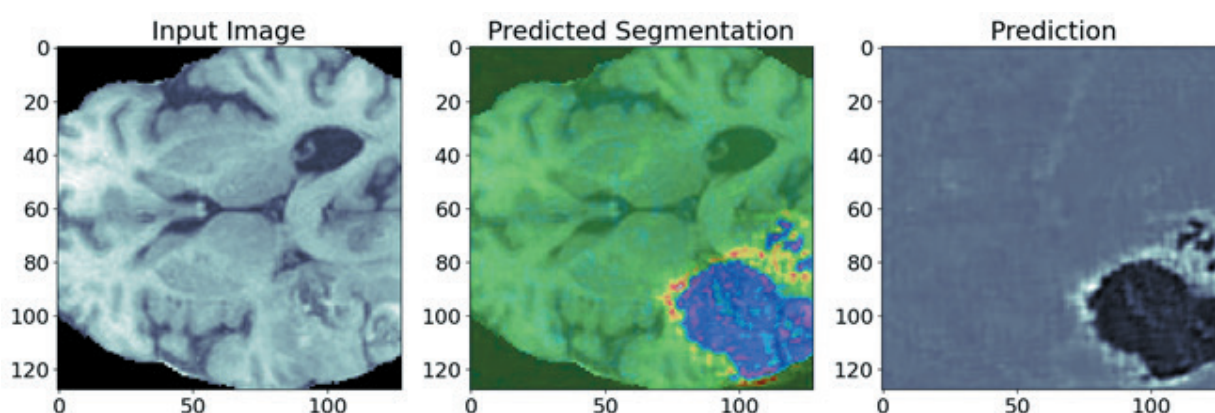


Figure 3. Visualization for the model's predicted segmentation sample



Hence, for tasks that resemble or are of the BraTS challenge, optimizers such as Adamax stand as strong candidates due to their ability to handle the complexities of medical image gradients. However, it is crucial to be mindful of its potential sensitivity to hyperparameter selection and conduct thorough tuning. AdamW remains a generally recommended optimizer for deep learning tasks, especially when dealing with complex datasets such as BraTS 2020, as well as models where regularization plays a vital role. The slight improvement observed over ordinary Adam further supports its use in medical image segmentation to achieve a good balance between performance and stability. Finally, future research should investigate the performance of these optimizers further across different BraTS datasets, as well as other models, potentially looking to hybrid models that utilize U-Net as a base model - potentially with more extensive hyperparameter tuning.

REFERENCES

- [1] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)", *IEEE Transactions on Medical Imaging* 34(10), 1993-2024 (2015) DOI: 10.1109/TMI.2014.2377694
- [2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", *Nature Scientific Data*, 4:170117 (2017) DOI: 10.1038/sdata.2017.117
- [3] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge", *arXiv preprint arXiv:1811.02629* (2018)
- [4] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al., "Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection", *The Cancer Imaging Archive*, 2017. DOI: 10.7937/K9/TCIA.2017.KLXWJJ1Q
- [5] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al., "Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection", *The Cancer Imaging Archive*, 2017. DOI: 10.7937/K9/TCIA.2017.GJQ7R0EF
- [6] Z. Zhang, "Improved Adam Optimizer for Deep Neural Networks," *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, Banff, AB, Canada, 2018, pp. 1-2, doi: 10.1109/IWQoS.2018.8624183.
- [7] J. -K. Fang, C. -M. Fong, P. Yang, C. -K. Hung, W. -L. Lu and C. -W. Chang, "AdaGrad Gradient Descent Method for AI Image Management," *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, Taoyuan, Taiwan, 2020, pp. 1-2, doi: 10.1109/ICCE-Taiwan49838.2020.9258085.
- [8] Huk, M. (2020). Stochastic Optimization of Contextual Neural Networks with RMSprop. In: Nguyen, N., Jearanaitanakij, K., Selamat, A., Trawiński, B., Chittayasothorn, S. (eds) *Intelligent Information and Database Systems. ACIIDS 2020. Lecture Notes in Computer Science()*, vol 12034. Springer, Cham.
- [9] Shazeer, Noam & Stern, Mitchell. (2018). Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. 10.48550/arXiv.1804.04235.
- [10] C. Arora, G. Raj, A. Ajit and A. Saxena, "ADAMAX-Based Optimization of Efficient Net V2 for NSFW Content Detection," *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Bangalore, India, 2023, pp. 1-6, doi: 10.1109/InC457730.2023.10263203.
- [11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
- [12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.



INTELLIGENT SOFTWARE SYSTEMS FOR MULTI-TENANT CLOUD ENVIRONMENTS: CHALLENGES AND SOLUTIONS

Veljko Pakević*,
[0009-0003-9932-4303]

Angelina Njeguš
[0000-0001-8682-7014]

Singidunum University,
Belgrade, Serbia

Abstract:

The rapid adoption of cloud computing has transformed software development and deployment, particularly through multi-tenant architectures that facilitate resource sharing while ensuring data isolation. This paper examines the role of intelligent software systems in multi-tenant cloud environments, addressing key challenges such as scalability, security, and adaptability to diverse user needs. A comprehensive analysis of existing research on multi-tenant cloud applications, security challenges, and IoT integration is presented. Additionally, the results of an experimental study on the use of autonomous drones for precision spraying in hazelnut plantations were observed and analysed. These drones operate within a cloud-based infrastructure hosted on Microsoft Azure, employing intelligent algorithms for route optimization, weather analysis, and real-time pesticide distribution adjustments. The multi-tenancy aspect is addressed by designing a system that enables multiple plantation owners to use a shared platform while maintaining individualized data and operational configurations. Our findings underscore critical concerns related to data privacy, performance optimization, and fault tolerance in shared cloud environments. Furthermore, we analyse security risks such as cross-tenant data leakage and access control vulnerabilities, offering recommendations for risk mitigation. This study contributes to the advancement of intelligent, cloud-based software architectures, showcasing how AI-driven systems can enhance efficiency and automation in multi-tenant applications. Future research will explore machine learning-based predictive analytics to further optimize resource utilization and strengthen security mechanisms in cloud-based IoT ecosystems.

Keywords:

Multi-Tenant Cloud, Intelligent Software Systems, IoT, Autonomous Drones, Cloud Security, AI in Agriculture.

INTRODUCTION

In the modern era, cloud computing serves as the foundation for the development and implementation of advanced software solutions. Multi-tenant cloud applications enable multiple users to simultaneously utilize a software system, where each user (tenant) has logically isolated data and personalized experience. This approach offers numerous advantages, including resource optimization, scalability, and reduced infrastructure maintenance costs. However, the implementation of intelligent software systems in a multitenant cloud environment presents several challenges, such as data security, system performance, adaptability to different users,

Correspondence:

Veljko Pakević

e-mail:

veljko.pakevic.24@singimail.rs





and efficient resource allocation. Additionally, the development and integration of intelligent algorithms in such environments require specialized strategies and architectural solutions to ensure optimal system performance.

The objective of this paper is to analyse the key aspects of applying intelligent software systems in multitenant cloud applications, highlighting challenges and proposed solutions. Understanding and enhancing intelligent software systems in multitenant cloud applications is a crucial step toward optimizing business processes, increasing security, and improving operational efficiency in environments that support multiple users. This paper aims to provide a comprehensive insight into the current state and potential future developments in this field, analysing concrete examples and trends shaping the future of intelligent software systems.

2. LITERATURE REVIEW

Cloud computing (CC) is a model enabling ubiquitous, convenient, on-demand access to a shared pool of configurable computing resources [1]. It is characterized by scalability, flexibility, and cost efficiency. Cloud environments are typically categorized into three main service models, such as: Infrastructure as a Service - IaaS (provides virtualized computing resources, such as virtual machines, storage, and networking), Platform as a Service - PaaS (development platform with tools and infrastructure for application deployment) and Software as a Service - SaaS (delivers software applications over the internet, eliminating the need for local installation and maintenance).

Multitenancy is a software architecture in which a single instance of software runs on a server and serves multiple tenants, ensuring data isolation and customization for each client within a shared infrastructure [2]. Each tenant operates within a logically isolated environment while sharing a common infrastructure, leading to cost savings and improved scalability. There are three primary multitenancy models:

1. Single Tenant: Each customer has a dedicated instance of the application and database.
2. Shared Database, Isolated Schema: A single database instance is used, but each tenant has a separate schema.
3. Shared Database, Shared Schema: All tenants share the same database schema, with mechanisms in place to differentiate their data.

Intelligent software systems integrate artificial intelligence (AI), specifically machine learning (ML) algorithms to enhance decision-making, automation, and predictive capabilities. These systems use algorithms to process data, recognize patterns, and adapt their behaviour. The key components of intelligent systems in cloud environments include:

1. Machine Learning Models: Algorithms that learn from data to make predictions and automate processes.
2. Natural Language Processing (NLP): Enables systems to understand and generate human language.
3. Predictive Analytics: Uses historical data to forecast future trends and behaviours.
4. Autonomous Decision-Making: AI-driven mechanisms that optimize system performance and resource allocation.

Intelligent software systems in multitenant cloud applications can be categorized based on their functionality and implementation into:

1. AI-Assisted Resource Management: uses ML models to predict and allocate cloud resources dynamically based on workload demands.
2. Security and Anomaly Detection: AI-driven systems analyse patterns in network traffic, user behaviour, and system logs to detect anomalies, help prevent security breaches, detect fraud, and respond to cyber threats in real time.
3. Personalized User Experience: AI models analyse user behaviour, preferences, and interactions to customize application UI, recommendations, and workflows, and enhance user engagement and retention through tailored experiences.
4. Automated Customer Support: AI-powered chatbots and virtual assistants handle common customer queries, reducing response times and improving efficiency.
5. AI-Powered Edge Computing & IoT Analytics: AI processes real-time IoT data at the edge to reduce cloud latency.
6. AI-Powered Workflow Automation: Automates repetitive tasks in cloud applications, improving productivity. Uses AI-driven decision-making for business process automation (BPA).



AI in multitenant cloud applications is expanding rapidly, enhancing efficiency, security, and user experience. From AI-driven resource management to intelligent automation, these systems provide robust solutions across multiple industries. However, several challenges and issues still need to be addressed. Some key issues are:

1. **Data Security and privacy concerns** - Multitenant cloud environments require strong data isolation to prevent leaks between tenants. AI-driven automation introduces additional security risks. Challenges are: data breaches due to misconfigured access controls, AI models unintentionally exposing sensitive tenant data, and compliance with GDPR, CCPA, HIPAA, and other data regulations.
2. **AI Bias & Ethical Concerns** - AI models may unintentionally favour certain groups due to biased training data. Challenges are AI-driven hiring, lending, and insurance decisions showing discrimination; lack of explainability in AI decision-making, and ethical concerns over AI surveillance and misuse of customer data.
3. **Scalability & Performance Bottlenecks** - AI-driven cloud applications require massive computing resources. Autoscaling isn't always efficient, therefore challenges are high latency in real-time applications, Cloud resource overutilization, leading to increased costs, and AI workload balancing across multiple tenants.
4. **Cost Management & Optimization** - AI-powered cloud solutions consume extensive resources, leading to high operational costs. Challenges are over-provisioning cloud resources in multitenant models, and unexpected cloud service cost spikes (e.g., AI training costs).
5. **Real-Time Data Processing Challenges** - Multitenant cloud applications generate massive real-time data streams from IoT, user activity, and AI workflows, therefore challenges are high-latency AI predictions for real-time applications and AI models struggling to process live sensor data efficiently.

Studies, described in Table 1, contribute valuable insights into the development and optimization of intelligent software systems within multitenant cloud environments, offering frameworks and methodologies to enhance security, performance, and resource management.

The study addresses the challenges inherent in multitenant cloud systems, such as data isolation risks, resource contention, and susceptibility to cyber threats. It explores the integration of adaptive resource management techniques and artificial intelligence (AI)--driven threat mitigation to enhance system robustness. The proposed framework utilizes machine learning models for dynamic resource allocation and real-time detection and prevention of cyber threats. The study found that integrating adaptive resource management with AI-driven threat mitigation significantly improved security and performance in multi-tenant cloud environments. The framework optimized workload distribution, reducing resource contention and enhancing overall system efficiency. AI-driven security mechanisms successfully identified and mitigated cyber threats in real-time, strengthening data protection. Performance metrics showed notable improvements, including reduced latency and increased throughput, while the risk of attacks was significantly minimized. Overall, the approach demonstrated a secure, scalable, and efficient solution for managing multi-tenant cloud computing systems. [3]

The paper provides a comprehensive analysis of the security challenges associated with deploying multitenant Field Programmable Gate Arrays (FPGAs) in cloud computing environments. It examines various deployment models, including spatial and temporal multi-tenancy, and evaluates their respective adversary models and security guarantees. The study highlights the vulnerabilities inherent in these models, particularly focusing on potential attacks such as Row hammer, cache side-channel attacks, and remote physical attacks that exploit the FPGA fabric. Additionally, the paper discusses the implications of untrusted cloud providers and the risks of intellectual property piracy, emphasising the need for robust security measures in multi-tenant FPGA deployments. The study found that integrating adaptive resource management with AI-driven threat mitigation significantly improved security and performance in multi-tenant cloud environments. The framework optimized workload distribution, reducing resource contention and enhancing overall system efficiency. AI-driven security mechanisms successfully identified and mitigated cyber threats in real-time, strengthening data protection. Performance metrics showed notable improvements, including reduced latency and increased throughput, while the risk of attacks was significantly minimized. Overall, the approach demonstrated a secure, scalable, and efficient solution for managing multi-tenant cloud computing systems. [4]



The research explores the integration of adaptive resource management and AI-driven threat mitigation to enhance security and performance in multi-tenant cloud computing environments. It addresses key challenges such as data isolation risks, resource contention, and cyber threats, proposing a framework that utilises machine learning algorithms for dynamic workload distribution and real-time threat detection. The study demonstrated that AI-powered security mechanisms significantly improved threat detection and prevention, reducing cyber risks. The adaptive resource management framework optimised workload distribution, leading to lower latency, increased system efficiency, and better scalability. The approach proved effective in securing multi-tenant cloud infrastructures while ensuring high performance and reliability. [5]

The study examines multi-tenant architectures in cloud computing, focusing on SaaS solutions, resource sharing, and data isolation. It explores AI-driven observability frameworks and advanced security mechanisms to enhance scalability and efficiency. Integrating AI and security measures improved scalability by 70% and reduced operational costs by 60%, optimising resource utilisation and system performance. [6]

Investigates the integration of AI into cloud security frameworks to enhance protection in multi-tenant environments. The study focusses on employing advanced threat detection and real-time monitoring mechanisms to identify and mitigate security threats proactively. The proposed model leverages ML algorithms to detect anomalous behaviours, predict potential security breaches and automate threat mitigation strategies. The study demonstrates that AI-driven security solutions significantly enhance threat detection accuracy and response times in multi-tenant cloud environments. The research highlights improvements in identifying unauthorised access and potential data breaches by implementing ML algorithms capable of real-time anomaly detection. Additionally, the paper addresses challenges related to integrating AI into existing cloud infrastructures, such as ensuring data privacy and managing computational overhead. [7]

The paper examines resilient multi-tenant cloud architectures, focusing on fault tolerance, security, and performance optimisation for reliable service. Implementing redundancy, data replication, and security protocols improved scalability, reliability, and data protection, ensuring uninterrupted cloud services. [8]

Provides a comprehensive analysis of security concerns associated with cloud computing (CC), focusing on various infrastructure levels: application, network, host, and data. It delves into significant challenges that could impact the CC business model and discusses documented solutions for each security issue. A particular emphasis is placed on multi-tenancy—a core feature of CC—which, while offering benefits like elasticity and flexibility, also introduces vulnerabilities such as abuse, unavailability, data loss, and privacy violations. The study aims to identify open problems and offers practical recommendations for future research to enhance overall CC security. The study underscores the substantial impact of multi-tenancy on cloud security, highlighting its contribution to various issues across all infrastructure levels. It emphasises the necessity for focused efforts to mitigate security vulnerabilities arising from multi-tenancy. By identifying specific challenges and suggesting targeted research directions, the paper contributes valuable insights to the broader discourse on cloud security, advocating for initiatives aimed at strengthening the resilience of cloud infrastructures. [9]

Additionally, since this paper examines this topic through an experimental study on the use of autonomous drones for precision spraying in plantations, where these drones operate within a cloud-based infrastructure employing intelligent algorithms for route optimisation, weather analysis, and real-time pesticide distribution adjustments, the analysis of relevant research papers is also analysed.

In a study [10], authors explored the use of reinforcement learning to optimise agricultural drone navigation for pesticide spraying. The research implements a Deep Q-learning algorithm that enables drones to autonomously adjust flight paths, avoid obstacles, and enhance spraying efficiency. The results demonstrate that this AI-driven approach significantly reduces pesticide wastage, minimises battery consumption, and improves overall spraying accuracy compared to traditional path-planning algorithms.

The research [11] focusses on optimising drone flight paths using a Bi-directional Long Short-Term Memory (Bi-LSTM) model integrated with a Deep Q-Network (DQN). This hybrid AI approach enhances UAV adaptability to changing environmental conditions while ensuring precise pesticide distribution. Experimental results demonstrate a reduction in redundant spraying, improved coverage accuracy, and lower operational costs for farmers.



The study [12] presents a data-driven optimization model for agricultural drone flight planning. By utilising a spray distribution model, the framework determines ideal flight speeds and pass widths to achieve uniform pesticide coverage. The study confirms improved spraying efficiency, minimised overlapping coverage, and better adaptability across different crop types and UAV models, enhancing precision agriculture practices.

The study [13] presents a modular, low-cost autonomous spraying control system integrated into traditional agricultural drones. The system utilises a machine learning-based spray uniformity algorithm, which optimises pesticide application by analysing real-time environmental data. By implementing a coefficient of variation-based approach, the study aims to ensure uniform pesticide distribution across plantations, reducing excessive chemical use and improving precision spraying efficiency. Experimental testing demonstrated a high level of spraying accuracy, with the system achieving a targeted distribution efficiency between 87.1% and 98.8%. The intelligent control mechanism significantly reduced over-spraying and under-spraying issues, leading to more effective pest control and lower chemical waste. The study confirms that AI-driven spray control can enhance agricultural productivity while minimising environmental impact.

3. METHODOLOGY AND EXPERIMENT

Traditional methods of crop spraying involve significant labour and machinery costs and inconsistent pesticide and fertiliser application. Optimising drone flight paths, ensuring precise pesticide application, and managing different plantation requirements are major challenges. Additionally, a scalable and intelligent system is needed to support multiple plantations while maintaining efficiency and security.

To address these challenges, we propose an intelligent, cloud-based drone management system that enables multiple plantation owners to utilise the same infrastructure while ensuring individual operational independence. The system leverages AI for flight path optimisation, real-time environmental monitoring, and automated pesticide application

System architecture and its components are:

- Multitenant Cloud Infrastructure (Azure-Based)
 - Each tenant operates within a logically isolated environment, allowing for personalized configurations and settings. This ensures that the operations and data of one tenant remain unaffected

by others.

- Azure Kubernetes Service (AKS) manages the deployment and scaling of containerized applications, such as drone control systems, ensuring high availability and efficient resource utilization.
- Azure IoT Hub facilitates secure and reliable communication between drones and the cloud, enabling real-time data collection and monitoring of drone operations.
- Azure SQL Database service stores plantation-specific data securely, implementing row-level security to ensure that each tenant can access only their respective data, maintaining data privacy and integrity.
- AI-Powered Flight Path Optimisation
 - Machine learning algorithms process terrain and weather data to plan optimal flight routes, considering factors like elevation changes and forecasted weather patterns.
 - Drones adjust their flight paths in real-time by analysing current wind speeds and humidity levels, ensuring stable flight and effective data collection.
- Intelligent Spraying System
 - Drones equipped with cameras and AI algorithms detect plant density and signs of disease, enabling targeted spraying only where needed.
 - Sensors monitor variables like wind direction and speed, ensuring chemicals are applied accurately, reducing drift, and enhancing effectiveness.
- Tenant Management and Access Control
 - Secure Login via Azure Active Directory B2C provides authentication and user management, allowing plantation owners to securely access their dedicated environments.
 - Implements strict access controls so that each tenant can only view and manage their data and analytics, preventing unauthorised access.
- Automated Reporting and Compliance
 - Azure Machine Learning generates predictive analytics on crop health and pest threats.
 - Reports are automatically generated for regulatory compliance and plantation management.

By integrating these components, the system aims to enhance the efficiency, precision, and sustainability of plantation management through advanced cloud-based solutions and artificial intelligence.



3.1. SCALABILITY

Challenge: During peak spraying seasons, the system must efficiently handle increased workloads without compromising performance.

Solution: Azure's auto-scaling capabilities dynamically adjust resources to meet demand. By utilising services like Azure App Service's automatic scaling, the system can automatically scale the number of application instances based on real-time metrics such as HTTP request volume or CPU usage. This ensures optimal performance during high-demand periods and cost savings during low-demand periods.

Implementation Steps

- Set rules that specify when to scale out (add resources) or scale in (remove resources) based on predefined thresholds.
- Use Azure Monitor to track application performance and trigger scaling actions as needed.
- Conduct load testing to ensure the system scales appropriately under various conditions.

3.2. DATA SECURITY

Challenge: Ensuring that each tenant's data remains secure and inaccessible to unauthorized users.

Solution: Implement Azure Role-Based Access Control (RBAC) and robust encryption mechanisms. RBAC allows you to assign specific permissions to users based on their roles, ensuring they access only the resources necessary for their tasks. Encryption safeguards data both at rest and in transit, protecting it from unauthorized access.

Implementation Steps

- Assign built-in or custom roles to users, groups, or service principals, limiting access to resources based on the principle of least privilege.
- Use Azure Disk Encryption for virtual machines and ensure that data stored in Azure Storage and Azure SQL Database is encrypted by default.
- Utilize Azure Key Vault to securely store and manage encryption keys, secrets, and certificates.

3.3. REAL-TIME DATA PROCESSING

Challenge: The system must process telemetry data from drones in real-time to dynamically adjust spraying parameters for optimal efficiency.

Solution: Leverage Azure Stream Analytics, a real-time analytics service designed to process large streams of data with low latency. By integrating Azure Stream Analytics with Azure IoT Hub, the system can ingest, process, and analyze telemetry data from drones, enabling immediate adjustments to spray parameters.

Implementation Steps:

- Configure Azure IoT Hub to collect telemetry data from drones.
- Develop queries that analyse incoming data streams and output results to appropriate services or storage.
- Use the analysed data to trigger functions or alerts that adjust drone operations in real-time.

3.4. COST MANAGEMENT

Challenge: Providing a cost-effective solution that scales with the needs of both small and large plantation owners.

Solution: Adopt Azure's pay-as-you-go pricing model, which charges users based on actual resource consumption. This model ensures that plantation owners pay only for the computing resources they use, making the system financially accessible regardless of farm size.

Implementation Steps:

- Utilise Azure Cost Management and Billing to track and analyse resource consumption.
- Establish spending limits and configure alerts to notify users when approaching budget thresholds.
- Regularly review and adjust resource allocations to prevent over-provisioning and minimise unnecessary costs.

The implementation of intelligent software systems in a multitenant cloud environment significantly enhances the efficiency of drone-based plantation spraying. By leveraging AI, IoT, and cloud computing, the proposed solution enables multiple plantation owners to share a robust, scalable, and intelligent infrastructure. This approach not only optimises spraying operations but also reduces costs, minimises environmental impact, and improves overall agricultural productivity.



4. CONCLUSION

Integrating intelligent software systems into multi-tenant cloud applications offers significant advantages and challenges. This study focused on multi-tenancy aspects such as resource optimisation, security, and system adaptability, particularly in agriculture, by examining autonomous drones for plantation spraying. Experiments demonstrated that a cloud-based, AI-driven drone management system on Microsoft Azure can effectively serve multiple plantation owners while ensuring data isolation and operational customisation. Results indicate that intelligent automation in agriculture enhances efficiency, reduces costs, and provides scalable solutions for precision farming. A review of existing literature on multi-tenant cloud computing, security, and IoT offered insights into technical and architectural challenges, including cross-tenant security risks, resource allocation, and performance optimisation. The study emphasises the need for robust isolation mechanisms, dynamic scaling, and AI-enhanced security solutions.

REFERENCES

- [1] S. K. a. S. Tim. Mather, "Cloud security and privacy: An Enterprise Perspective on Risks and Compliance, First. Sebastopol," in *O'Reilly*, 2009.
- [2] N. Ruparelia, "Cloud Computing, Revised and Updated Edition," 2023. [Online]. Available: <https://mitpress.mit.edu/9780262546478/cloud-computing>. [Accessed 1 April 2025].
- [3] M. Sathik, "Enhancing Security and Performance in Multi-Tenant Cloud Computing Environments Through Adaptive Resource Management and AI-Driven Threat Mitigation - QIT Press," 3 2025. [Online]. Available: https://qitpress.com/articles/QITP-IJCC_05_01_002. [Accessed 27 Mart 2025].
- [4] G. Dessouky, A. R. Sadeghi and S. Zeitouni, "SoK: Secure FPGA multi-tenancy in the cloud: Challenges and opportunities," in *2021 IEEE European Symposium on Security and Privacy, Euro S and P 2021*, 2021.
- [5] M. R. Mekala, "AI-driven optimization for multi-tenant cloud platforms: balancing cost, performance, and security," *International journal of computer engineering and technology*, vol. 16, no. 1, pp. 1381-1400, 1 2025, doi: https://doi.org/10.34218/IJCET_16_01_104
- [6] R. K. Sharma, "Multi-Tenant Architectures in Modern Cloud Computing: A Technical Deep Dive," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 1, pp. 307-317, 1 2025, doi: <https://doi.org/10.32628/CSEIT25111236>.
- [7] S. Chippagiri, "A Study of Cloud Security Frameworks for Safeguarding Multi-Tenant Cloud Architectures," *International Journal of Computer Applications*, vol. 186, no. 60, pp. 50-57, 2025.
- [8] S. A. Dave, N. K. Gannamneni, B. Gajbhiye, R. Agarwa, S. Jain and P. K. Gopalakrishna, "Designing Resilient Multi-Tenant Architectures in Cloud Environments," *International Journal for Research Publication and Seminar*, vol. 11, no. 4, pp. 356-373, 12 2020, doi: <https://doi.org/10.36676/JRPS.V11.14.1586>.
- [9] M. A. Hayat, S. Islam and M. F. Hossain, "Securing the Cloud Infrastructure: Investigating Multi-tenancy Challenges, Modern Solutions and Future Research Opportunities," *International Journal of Information Technology and Computer Science*, vol. 16, no. 4, pp. 1-28, 2024.
- [10] Y.-Y. Huang, Z.-W. Li, C.-H. Yang and Y.-M. Huang, "Automatic Path Planning for Spraying Drones Based on Deep Q-Learning," *Journal of Internet Technology*, vol. 24, no. 3, pp. 565-575, 2023.
- [11] H. Fu, Z. Li, W. Zhang, Y. Feng, L. Zhu, X. Fang and J. Li, "Research on Path Planning of Agricultural UAV Based on Improved Deep Reinforcement Learning," *Agronomy*, vol. 14, no. 11, p. 2669, 2024.
- [12] R. V. Nanavati, Y. Meng, M. Coombes and C. Liu, "Generalized data-driven optimal path planning framework for uniform coverage missions using crop spraying UAVs," *Precision Agriculture*, vol. 24, no. 4, pp. 1497-1525, 2023.
- [13] P. Wang, A. S. Hanif, S.-H. Yu, C.-G. Lee, Y. H. Kang, D.-H. Lee and X. Han, "Development of an autonomous drone spraying control system based on the coefficient of variation of spray distribution," *Computers and Electronics in Agriculture*, vol. 227, no. 1, 2024.



SPEED CONTROL OF A ROTARY SERVO-BASE UNIT: LYAPUNOV AND MIT RULE APPROACHES

Mitra Vesović*,
[0000-0003-0457-1874]

Vladimir Zarić,
[0000-0002-0999-0608]

Natalija Perišić,
[0000-0002-8675-1934]

Radiša Jovanović
[0000-0002-8122-756X]

Faculty of Mechanical Engineering,
Belgrade, Srebia

Abstract:

This paper presents a comparative study of two speed control techniques for a rotary servo-base unit using Lyapunov-based adaptive control and the MIT rule technique. The primary objective is to achieve precise and stable speed control and to analyze the influence of adaptation gain on the system performance. A mathematical model of the rotary system is analyzed, followed by the development of adaptive controllers based on the Lyapunov stability theory and the MIT rule. Choosing a suitable reference model is examined, and parameter adaptation laws are designed to optimize system performance. The impact of different adaptation gains on system response is evaluated through simulations in MATLAB/Simulink. Figures illustrating the evolution of adaptation parameters over time, as well as system response, are provided. Various performance criteria, settling time, overshoot, and different objective functions are used to compare the control approaches. The results highlight the advantages and limitations of each method. Recommendations for tuning adaptation parameters are provided to improve overall system performance.

Keywords:

Lyapunov Rule, MIT Rule, DC Motor, Model Reference Adaptive Control (MRAC), Adaptation Gain.

INTRODUCTION

Model reference adaptive control (MRAC) is an adaptive control strategy that develops a control law using an adjustable gain, making the system's plant continuously follow a reference model until the tracking error becomes zero [1]. The Massachusetts Institute of Technology (MIT) rule adjusts controller parameters using the gradient method based on the error between the plant's and reference model's output. However, a system created using the MIT rule can sometimes become unstable. In contrast, the Lyapunov approach guarantees stability by using the Lyapunov function that depends on the output and parameter error, ensuring the system remains stable as long as the derivative of the Lyapunov function is negative. Lyapunov and MIT rule approaches were compared in control of the coupled tank systems in the MRAC scheme in [2], or for the control of similar systems in [1], where fuzzy-optimized MRAC was applied.

Correspondence:

Mitra Vesović

e-mail:

mvesovic@mas.bg.ac.rs



In this study, adaptive controllers are developed based on Lyapunov and MIT rules for speed control of DC motor, and a comparison between these two approaches is made.

2. DESCRIPTION OF THE SYSTEM

The creation of a mathematical model is among the first stages in the development of a control system. This saves time and profit in the long run [3]. In this section, we investigate the dynamics of a DC motor, which serves as the object in our control system. A full schematic representation of our object is shown in Figure 1.

2.1. ELECTRICAL DYNAMICS: VOLTAGE AND CURRENT EQUATIONS

The equations specifying the motor's electrical assemblies are listed below in Equation 1:

$$V_m(t) = R_m I_m(t) + L_m \frac{d}{dt} I_m(t) + e_b(t) \quad (1)$$

$$e_b(t) = k_m \omega_m(t) \quad (2)$$

Equation 1. Electrical dynamics of the object

Here: V_m , e_b , k_m , and ω_m are motor voltage, back electromotive voltage, back electromotive voltage constant, and speed of the motor shaft, respectively. Since the motor inductance L_m is much less than its resistance R_m , it can be ignored [4]. Solving the system of equations for motor current I_m , we get an electrical equation of a DC motor.

$$I_m(t) = \frac{V_m(t) - k_m \omega_m(t)}{R_m}$$

Equation 2. DC motor current

2.2. MECHANICAL DYNAMICS: THE SECOND NEWTON'S LAW OF MOTION

We consider the second Newton's law of motion and relationships between the following quantities: moment of inertia of the load J_p of the motor shaft J_m ; speed of the load shaft ω_p ; viscous friction on both motor shaft B_m and the load shaft B_p . Total torques applied on the load τ_l and on the motor τ_m , with resulting torque acting on the motor shaft from the load torque denoted as τ_{ml} , are given with the following Equation 3:

$$J_l \frac{d\omega_l(t)}{dt} + B_l \omega_l(t) = \tau_l(t) \quad (1)$$

$$J_m \frac{d\omega_m(t)}{dt} + B_m \omega_m(t) + \tau_{ml}(t) = \tau_m(t) \quad (2)$$

Equation 3. Torque and motion analysis

Equation 4 represents the mechanical dynamics of the rotary servo base unit, which can be calculated with J_{eq} and B_{eq} as the total moment of inertia and damping term using the object's constants: η_g and K_g , which are the gearbox efficiency and the total gear ratio, respectively.

$$J_{eq} \frac{d\omega_l(t)}{dt} + B_{eq} \omega_l(t) = \eta_g K_g \tau_m(t)$$

Equation 4. Mechanical dynamics of the object

2.3. SYSTEM DYNAMICS INTEGRATION

Finally, when assuming that motor torque is proportional to voltage and with the introduction of the equivalent damping term $B_{eq,v}$, and the actuator gain A_m , consisting of motor efficiency η_m and current-torque constant k_p , the electromechanical model, Equation 5, is as follows:

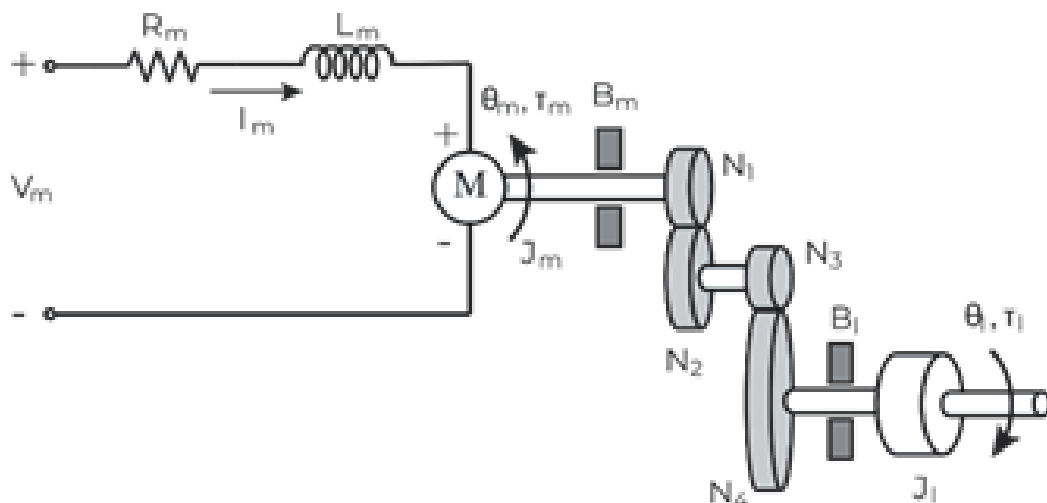


Figure 1. Illustration of the object under study



$$J_{eq} \left(\frac{d}{dt} \omega_l(t) \right) + B_{eq,v} \omega_l(t) = A_m V_m(t) \quad (1)$$

$$B_{eq,v} = \frac{\eta_g K_g^2 \eta_m k_t k_m + B_{eq} R_m}{R_m} \quad (2)$$

$$A_m = \frac{\eta_g K_g \eta_m k_t}{R_m} \quad (3)$$

Equation 5. Unified system representation

Choosing motor voltage as input $V_m(t)=u$, and angular speed of the load shaft as the output variable, $\omega_l(t)=y$ the system is now defined with $J_{eq}=0.0021 \text{ kgm}^2$, $B_{eq,v}=0.084 \text{ kgm}^2 \text{ s}^{-1}$ and $A_m=0.1284 \text{ Nm/V}$.

$$\dot{y} = -\frac{B_{eq,v}}{J_{eq}} y + \frac{A_m}{J_{eq}} u$$

Equation 6. State equation of the object

The system's block diagram is given in Figure 2.

3. MODEL REFERENCE ADAPTIVE CONTROL (MRAC)

Figure 3 displays the structural diagram of a typical MRAC system. It has two loops: one for parameter modification and one for feedback. In this diagram, y_d is the set point or reference, y_m is the reference model's output, y is the object's output, and u is the input (and control law which depends on the adaptation parameters) in the object.

To adjust the control algorithm and make the object track the reference model's output, y_m , the controller's parameters are tuned. The adaptation law or adjustment mechanism of the MRAC system can be found using a variety of methods. The gradient approach, also known as the MIT rule, a stability theory the Lyapunov method, or some others can all be used to carry out the MRAC adjustment mechanism [5].

3.1. LYAPUNOV RULE

Since the plant of interest in this paper is the first-order object, Equation 6, we will consider a first-order plant and a reference model given by all positive coefficients (a, b, a_m, b_m) and control algorithm.

$$\frac{dy}{dt} = -ay + bu \quad (1)$$

$$\frac{dy_m}{dt} = -a_m y_m + b_m y_d \quad (2)$$

$$u = \theta_1 y_d - \theta_2 y \quad (3)$$

Equation 7. First-order adaptive control structure

When we substitute Equation 7(3) into Equation 7(1) and take the derivative of the error function, which we define as the difference between the real output of the plant, y , and the output of the reference model, y_m , we obtain:

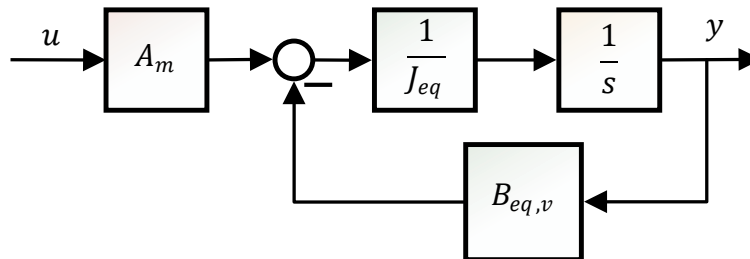


Figure 2. Block diagram of a linear system

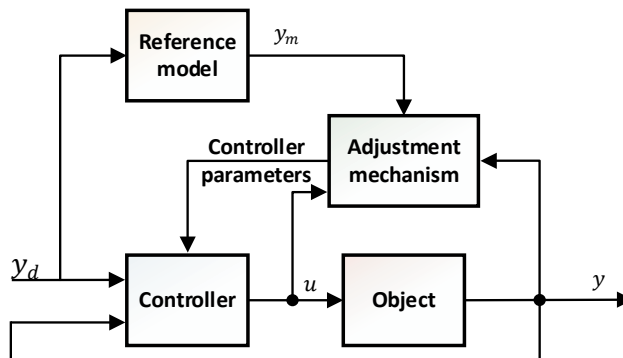


Figure 3. Structural diagram of a general idea for MRAC

$$e = y - y_m \quad (1)$$

$$\dot{e} = -a_m e - (b\theta_2 + a - a_m)y + (b\theta_1 - b_m)y_d \quad (2)$$

Equation 8. The error function and its derivation

The algorithms for changing the parameters in the MRAC system can be explained by the Lyapunov stability theory. A Lyapunov function, V , is necessary for the Lyapunov method. It must be positive definite, and its derivative, $\dot{V}$, must be negative definite. Since there is no methodical approach to determining an appropriate Lyapunov function [6], the Lyapunov function and its derivative are chosen as:

$$V(e, \theta_1, \theta_2) = \frac{1}{2} \left(e^2 + \frac{1}{b\gamma} (b\theta_2 + a - a_m)^2 + \frac{1}{b\gamma} (b\theta_1 - b_m)^2 \right) \quad (1)$$

$$\dot{V} = -a_m e^2 + \frac{1}{\gamma} (b\theta_2 + a - a_m)(\dot{\theta}_2 - \gamma y e) + \dots \quad (2)$$

$$\dots + \frac{1}{\gamma} (b\theta_1 - b_m)(\dot{\theta}_1 + \gamma y_d e)$$

Equation 9. Lyapunov function with derivative

with positive γ . For $\dot{V}$ to be negative definite, we will cancel the second and third terms in Equation 9(2) by ensuring that the adjustable parameters are updated as:

$$\frac{d\theta_1}{dt} = -\gamma y_d e \quad (1)$$

$$\frac{d\theta_2}{dt} = \gamma y e \quad (2)$$

Equation 10. Lyapunov rule adaptation

where γ represents the tuning parameter (or adaptation parameter, adaptation gain). These two equations from Equation 10 are Lyapunov adjusting mechanisms I and II. They are shown graphically in Figure 4.

3.2. THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY (MIT) RULE

Derivation of the MIT rule is also achieved using the error function from Equation 8 and objective function $J(\theta) = \frac{1}{2} e^2$ with its derivative $\frac{\partial J}{\partial e} = e$. The goal is adjusting the parameter $\theta = [\theta_1 \ \theta_2]$ so that direction of the negative gradient of the objective function is guaranteed with the adaptation gain γ_1 , which is utilized to modify the controller's adaption rate: $\frac{d\theta}{dt} = -\gamma^1 \frac{\partial J}{\partial \theta} = -\gamma^1 e \frac{\partial e}{\partial \theta} \rightarrow \frac{d\theta_1}{dt} = -\gamma^1 e \frac{\partial e}{\partial \theta_1}$ and $\frac{d\theta_2}{dt} = -\gamma^1 e \frac{\partial e}{\partial \theta_2}$. Similar to the first step in the previous section, when we insert the third equation of Equation 7 into the first and compare the actual and reference models, taking into account that $y_m = y$ (or at least asymptotically trace), we get:

$$\theta_1 = \frac{b_m}{b} \quad (1)$$

$$\theta_2 = \frac{a_m - a}{b} \quad (2)$$

Equation 11. Control parameters in MIT rule using reference model

Taking the Laplace transform of the first two equations of Equation 7 (with substituting the third equation into the first) and inserting them into the definition of the error, Equation 8(1), the following equations are obtained:

$$e(s) = \frac{b\theta_1}{s+a+b\theta_2} y_d - \frac{b_m}{s+a_m} y_d \quad (1)$$

$$\frac{\partial e}{\partial \theta_1} = \frac{b}{s+a+b\theta_2} y_d \stackrel{\text{Equation 11(2)}}{=} \frac{b}{s+a_m} y_d \quad (2)$$

$$\frac{\partial e}{\partial \theta_2} = -\frac{b^2 \theta_1}{(s+a+b\theta_2)^2} y_d \stackrel{\text{Equation 11(2)}}{=} -\frac{b}{s+a_m} y \quad (3)$$

Equation 12. Partial error derivation

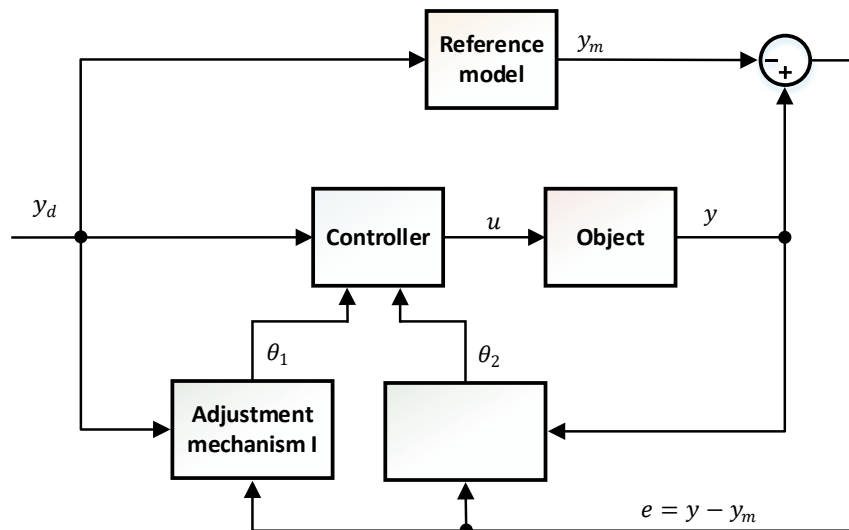


Figure 4. MRAC structural diagram using the Lyapunov rule



Finally, the MIT rule-based adaptation laws include:

$$\frac{d\theta_1}{dt} = -\gamma e \frac{a_m}{s+a_m} y_d \quad (1)$$

$$\frac{d\theta_2}{dt} = \gamma e \frac{a_m}{s+a_m} y \quad (2)$$

Equation 13. MIT rule adaptation,

where $\gamma = \frac{\gamma^1 b}{a_m}$. It is found that the adaptation laws derived from the Lyapunov and MIT rules are alike (see Figure 5), with the exception that the MIT rule adds a filter that is equal to the reference model's transfer function [7].

4. RESULTS

The reference model used in this paper is in the very simple transfer form: $W_{ref} = \frac{1}{0.02s+1}$. In Figure 6, results for the reference speed of $y_d = 1 \frac{rad}{s}$ are presented for different adaptation gains γ . The adaptive gain impacts stability and convergence, with larger gains causing faster response times but potentially large overshoots and oscillations, while smaller values improve stability but slow the system response.

In general, the properties of the control system and the performance requirements determine which adaptation gain should be used.

Comparison and the system's dynamic behavior quality for both rules are given in Table 1 and in Figure 7. Since responses for $\gamma = 0.01$ and $\gamma = 0.1$ do not reach the reference in 4s, the overshoot and the settling time ($\pm 2\%$ of the steady-state value) are not calculated for them. Four different objective functions are calculated: Integral Square Error (ISE), Integral Absolute Error (IAE), Integral Time-weighted Absolute Error (ITAE), and Integral Time-weighted Square Error (ITSE). Results show, both in Table 1 and Figure 7, that higher values of the adaptation parameter guarantee smaller errors with shorter settling times.

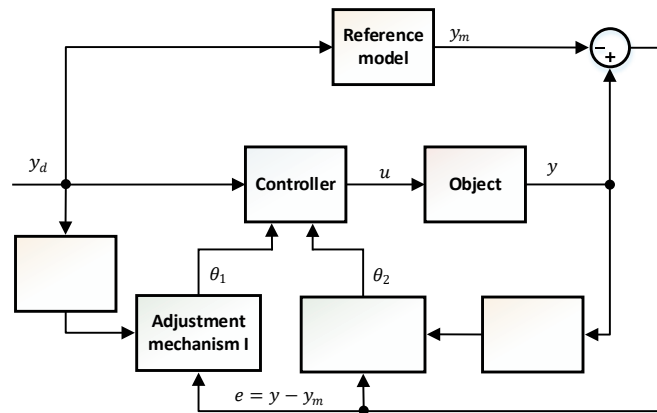


Figure 5. MRAC structural diagram using the MIT rule

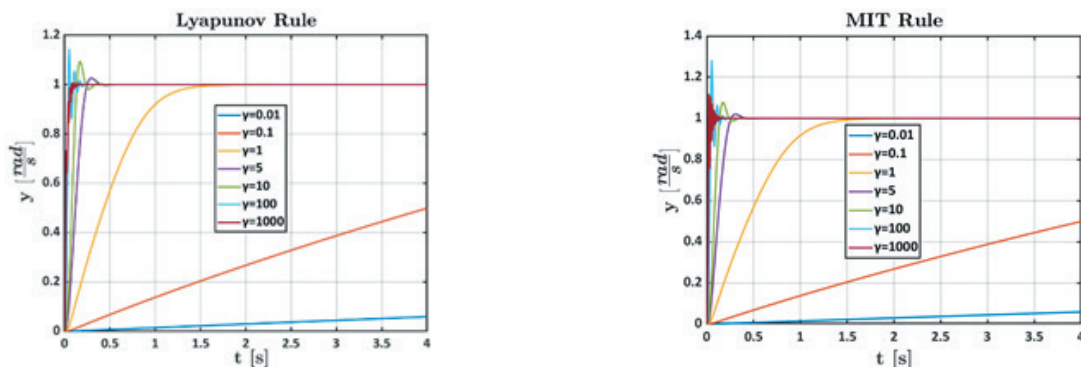


Figure 6. Simulation results for the Lyapunov rule (left) and MIT rule (right)



Table 1. Comparison and measurement quality

Lyapunov rule								MIT rule				
Adaptation gain	Overshoot [%]	Settling time [s]	Objective functions				Overshoot [%]	Settling time [s]	Objective functions			
			ISE	IAE	ITAE	ITSE			ISE	IAE	ITAE	ITSE
0.01	/	/	0.4668	0.4784	0.1240	0.1232	/	/	0.4788	0.4864	0.1243	0.1236
0.1	/	/	0.4513	0.4723	0.1194	0.1142	/	/	0.4513	0.4723	0.1194	0.1142
1	0	1.2979	0.2470	0.3413	0.0772	0.0500	0	1.3098	0.2580	0.3489	0.0774	0.0503
5	2.59	0.3273	0.0607	0.1050	0.0096	0.0044	1.97	0.2445	0.0715	0.1132	0.0098	0.0048
10	9.30	0.2183	0.0327	0.0647	0.0045	0.0016	7.74	0.2202	0.0427	0.0715	0.0045	0.0019
100	14.40	0.1456	0.0042	0.0183	8.7E-4	1.1E-4	28.24	0.1478	0.0098	0.0263	0.0011	2.4E-4
1000	0.78	0.0879	4.1E-4	0.0051	1.9E-4	1.0E-5	11.89	0.0988	0.0021	0.0105	3.4E-4	4.0E-5

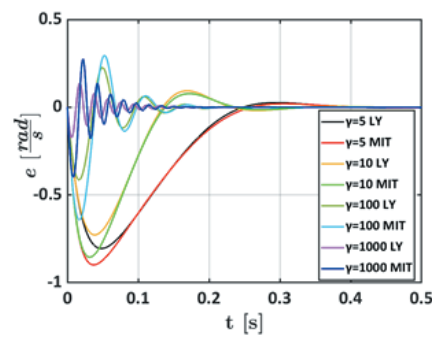
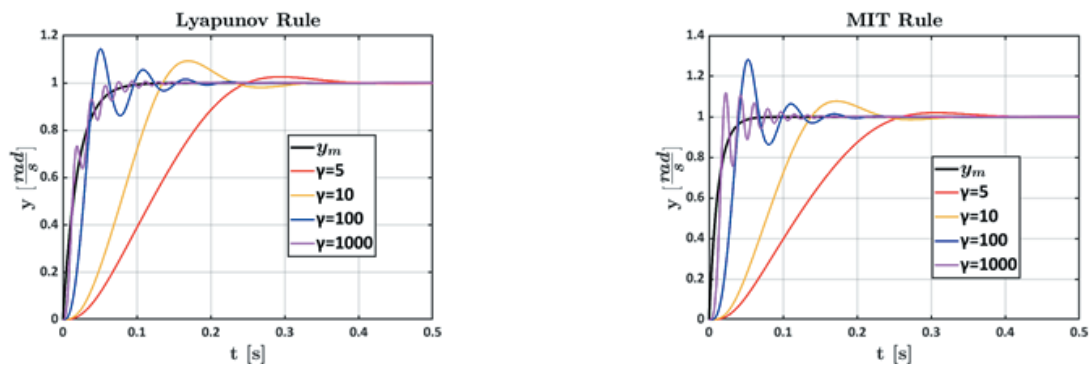
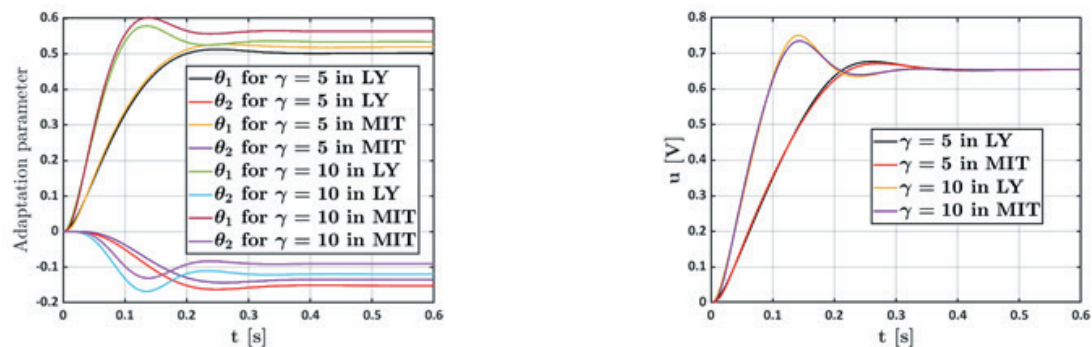
Figure 7. Error signal for Lyapunov and MIT rules for different γ values

Figure 8. Difference between reference model and plant output for Lyapunov (left) and MIT rule (right)

Figure 9. Changes in the parameters of adaptation θ_1 and θ_2 (left) and in the control signals (right)



The following Figure 8 shows the difference between output reference model y_m and plant output y for the Lyapunov and MIT rules only for $\gamma > 1$. Changes in the Lyapunov (Equation 10) and in MIT (Equation 13) rule adaptation are shown in the following Figure 9 (left) for two different tuning parameters: $\gamma = 5$ and $\gamma = 10$. For the same tuning parameters, Figure 9 (right) shows the control signal.

For further testing, a special function was designed to generate different reference motor speeds. The function is given in the form of the time-dependent piecewise constant function.

Figure 10 shows the response of the system for the reference speed of the mentioned constant piecewise function as the reference output for both the Lyapunov and MIT rules. That higher values of the adaptive gain proved to be a better choice is confirmed also in Figure 11, where the sinusoidal signal was selected as the reference. For both the Lyapunov and MIT techniques, figures for system response and control signals appear nearly identical, with differences not explicitly visible.

5. CONCLUSION

In this study, the Lyapunov and MIT adaptive control algorithms were investigated and applied to control the speed of a rotary servo-base unit. A first-order transfer function was created as the reference model to ensure a smooth system response and match with the dynamics of the real object. The time constant of the chosen reference model is 0.02s, which enables the DC motor to track the reference signal with a minimal delay while avoiding oscillations or overshoots. The trial-and-error approach was applied to determine the adaptation gain γ , leading to the conclusion that higher values (above 1) result in improved system performance. This paper offers practical guidelines for selecting the reference model and the adaptation gain in model reference adaptive control applications. For future research, more advanced approaches, such as machine learning methods and fuzzy-neural or metaheuristic optimization algorithms, could be employed to find out the reference model and adaptation parameters in the control system.

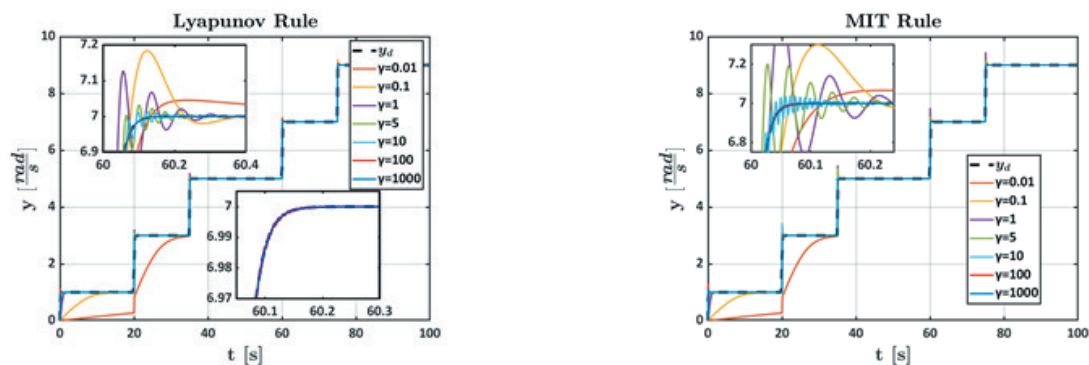


Figure 10. Simulation results for step changeable reference signal

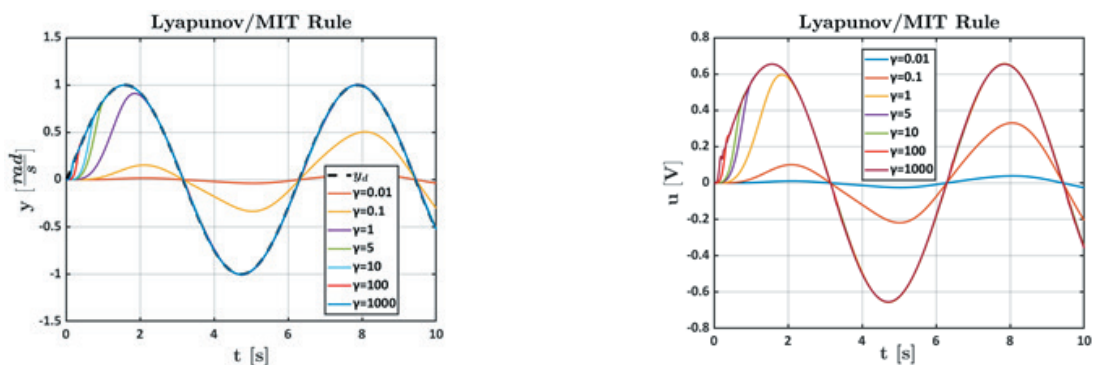


Figure 11. Simulation results for a sinusoidal reference signal (left) and control signal (right)



6. ACKNOWLEDGEMENTS

This work was financially supported by the Ministry of Education, Science and Technological Development of the Serbian Government, under contract 451-03-137/2025-03/200105, from date 04.02.2025.

REFERENCES

- [1] D. D. Dinakin and P. O. Oluseyi, "Fuzzy-Optimized Model Reference Adaptive Control of Interacting and Noninteracting Processes Based on MIT and Lyapunov Rules," *Turkish Journal of Engineering*, vol. 5, no. 4, pp. 141-153, 2021.
- [2] D. Gupta, A. Kumar, and V. Kumar Giri, "Investigations on MIT and Lyapunov Rule-Based Modified MRAC for Noninteracting and Interacting Two-Tank Coupled Systems," *FME Transactions*, vol. 52, no. 2, pp. 305-318, 2024.
- [3] M. Vesović and R. Jovanović, "Grey Wolf Optimization for Position Control of a Direct Current Motor Driven by Feedback Linearization Method," in *Sinteza 2022 - International Scientific Conference on Information Technology and Data Related Research*, Belgrade, 2022.
- [4] J. Apkarian, M. Levis, and H. Gurocak, *SRV02 Base Unit Experiment for MATLAB/Simulink Users*, Markham: Quanser Inc., 2012.
- [5] K. J. Astrom and B. Wittenmark, "Adaptive Control, 2nd ed.," Pearson Education Asia, 2001, pp. 185-225.
- [6] K. J. Astrom and B. Wittenmark, *Adaptive Control*, USA: Addison-Wesley, 1989.
- [7] P. Swarnkar, S. K. Jain, and R. K. Nema, "Comparative Analysis of MIT Rule and Lyapunov Rule in Model Reference Adaptive Control Scheme," *Inovative Systems Design and Engineering*, vol. 2, no. 4, pp. 154-162, 2011.



CONTROL OF A LIQUID LEVEL SYSTEM BASED ON A PROPORTIONAL-SUM CONTROLLER USING WHALE OPTIMIZER

Vladimir Zarić*,
[0000-0002-0999-0608]

Natalija Perišić,
[0000-0002-8675-1934]

Mitra Vesović,
[0000-0003-0457-1874]

Radiša Jovanović
[0000-0002-8122-756X]

Faculty of Mechanical Engineering,
Belgrade, Srebia

Abstract:

This study is about controlling the liquid level in water tanks, which is one of the challenges in the industry. A nonlinear analytical model is derived and validated against linearized continuous-time and discrete-time models, demonstrating their equivalence under nominal conditions. The classical method for tuning a discrete-time proportional-sum controller was applied, with gains tuned via the Ziegler-Nichols method. Additionally, the controller's parameters were fine-tuned by using the whale optimization algorithm. Simulation results for the tank system are presented. While Ziegler-Nichols gives a decent base, it can be said that optimization of the controller parameters should be recommended when dealing with similar problems in real-life situations. Results reveal that the optimized controller reduces the sum of squared errors compared to the classical controller, achieving superior accuracy.

Keywords:

Discrete-Time Systems, Whale Optimization Algorithm, Proportional-Difference-Sum Controller, Liquid Level Control, Ziegler-Nichols method.

INTRODUCTION

Control of liquid levels in tanks is always an important research topic, and a lot of scientific work has been done in order to find an optimal solution for this problem. The reason for this constant interest lies in the fact that tank plants can be found in multiple industries, like as chemical, pharmaceutical, oil and gas, water treatment, and so on.

When it comes to the utilization of PID-like controllers or their discrete-time equivalent proportional-difference-sum (PDS) controllers, the main task is to obtain PID parameters that provide good system behavior. Choosing proper parameters can significantly improve the performance of the system, while poor tuning can worsen it [1]. Authors in [1] offer a few conventional methods for tuning a PID controller. Although those methods may achieve the desired performance, the authors stated that a lot of effort and experience are needed for defining parameters. Intelligent controllers can also be used for this purpose.

Correspondence:

Vladimir Zarić

e-mail:

vzaric@mas.bg.ac.rs





A classical fuzzy logic controller is implemented and compared to the PID controller in terms of obtaining the desired system response in [2]. Their result gave an advantage to the fuzzy controller. However, authors in paper [3] state in favor of PID controllers when it comes to their simple structure and easy tuning, and recommend using them for liquid level plants without substantial change in their dynamics. To regulate liquid levels under time-varying system behavior across diverse operational regimes, the authors employed a fuzzy-logic-based adaptive PID control scheme with gain scheduling. Artificial neural networks also found their application for solving this task. Paper [4] offers two different liquid level control strategies based on neural networks. The initial approach employs an inverse-model-driven neurocontroller, whereas the alternative strategy utilizes a neural-network-enhanced predictive control scheme. Reference [5] combines neural networks and fuzzy logic to propose a self-tuning neuro-fuzzy regulator designed for liquid level regulation, and it outperformed the fuzzy logic controller. The fusion of AI-based methods and conventional PID control can be achieved through evolutionary optimization strategies to adjust the PID gains. For example, in [6], for tuning PID controller parameters Genetic algorithm is used, and a comparison is made with PID tuned via the ZN method. As it was expected, including the metaheuristic algorithm brought better results. Sometimes, researchers modify the origi-

nal metaheuristic algorithm, like in [7], where the Grey Wolf Optimization algorithm is modified for tuning PID. The increasing use of optimization algorithms in controlling systems is also reflected in their application in system modeling, like in [8], where the whale optimization algorithm (WOA) is employed to obtain an optimized Takagi-Sugeno plant model. Optimization algorithms are applicable in areas other than automatic control, such as in medicine [9], or in economics [10], where authors have used modified WOA to solve various tasks.

This work demonstrates liquid level control using a conventional proportional-sum (PS) controller whose parameters were optimized by the WOA [11]. The results are compared with the PS controller, whose parameters are obtained using the classical ZN tuning method [12].

2. SYSTEM CHARACTERIZATION

The system's physical characteristics employed in this study are presented in Table 1.

The water pump, reservoir, and two similar cylinder tanks, one above the other, make up the system. Water is pumped vertically from the reservoir to the upper tank through a pumping system. We need to control the water level of the second tank. The used system diagram is shown in Figure 1.

Table 1. Parameter configuration

Parts of the system	Labels	Numerical values	Unit of measurement
Pump flow constant	K_p	$5.37 \cdot 10^{-6}$	$m^3s^{-1}V^{-1}$
Diameter of outlet opening 1	D_{o1}	$0.47625 \cdot 10^{-2}$	m
Diameter of outlet opening 2	D_{o2}	$0.47625 \cdot 10^{-2}$	m
Tank 1 inside diameter	D_1	$4.445 \cdot 10^{-2}$	m
Tank 2 inside diameter	D_2	$4.445 \cdot 10^{-2}$	m
Gravitational constant	g	9.81	ms^{-2}

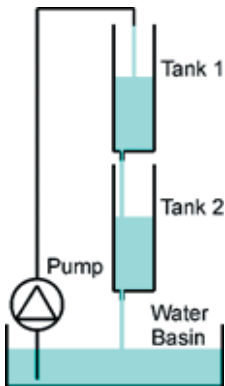


Figure 1. Liquid level system



3. MATHEMATICAL MODELING OF THE SYSTEM

3.1. ANALYTICAL NONLINEAR MODEL

Tank 1 and Tank 2 are the two subsystems of the plant. Pump voltage V_p is the input in tank 1, and tank 1 water level H_1 is the output. One way to express the flow into tank 1 is

$$Q_{i1} = K_p V_p.$$

Equation 1. Flow into tank 1

The outflow velocity V_{o1} and the opening cross-sectional area of tank 1, A_{o1} , are multiplied to give the outflow from tank 1,

$$Q_{o1} = A_{o1} V_{o1}.$$

Equation 2. Outflow from tank 1

The mass balance equation for tank 1 is

$$A_1 \frac{dH_1}{dt} = Q_{i1} - Q_{o1} = K_p V_p - A_{o1} \sqrt{2gH_1},$$

Equation 3. Mass balance equation for tank 1

where A_1 is the cross-sectional area of tank 1. Outflow velocity V_{o2} and the opening cross-sectional area of tank 2, A_{o2} , are multiplied to give the outflow from tank 2,

$$Q_{o2} = A_{o2} V_{o2}.$$

Equation 4. Outflow from tank 2

The mass conservation equation governing tank 2 is expressed as:

$$A_2 \frac{dH_2}{dt} = Q_{i2} - Q_{o2} = A_{o1} \sqrt{2gH_1} - A_{o2} \sqrt{2gH_2}.$$

Equation 5. Mass balance equation for tank 2

3.2. LINEAR MODELS

Since the water level in reservoir 2 is supposed to have a constant nominal value for steady-state conditions, the water level in reservoir 1 and the pump voltage also have constant values:

$$H_1 = H_{1N}, H_2 = H_{2N}, V_p = V_{pN}.$$

Equation 6. Nominal values for steady-state conditions

The next action to take is to use Taylor's series representation at nominal values, Equation 6, to approximate nonlinear functions given in Equation 7,

$$f_1 = \frac{dH_1}{dt} = \frac{K_p V_p}{A_1} - \frac{A_{o1} \sqrt{2gH_1}}{A_1}, \quad (1)$$

$$f_2 = \frac{dH_2}{dt} = \frac{A_{o1} \sqrt{2gH_1}}{A_2} - \frac{A_{o2} \sqrt{2gH_2}}{A_2}. \quad (2)$$

Equation 7. Nonlinear functions

As a result, the following linear differential equations are obtained:

$$\dot{h}_1 = a_1 h_1 + b_1 v_p, \quad (1)$$

$$\dot{h}_2 = a_2 h_2 + b_2 h_1. \quad (2)$$

Equation 8. Linear differential equations

Variables h_1 , h_2 , and v_p in Equation 8 stand for deviations from nominal values:

$$h_1 = H_1 - H_{1N}, \quad (1)$$

$$h_2 = H_2 - H_{2N}, \quad (2)$$

$$v_p = V_p - V_{pN}, \quad (3)$$

Equation 9. Deviations from nominal values

and coefficients a_1 , a_2 , b_1 and b_2 are calculated using the expressions below,

$$a_1 = -\frac{A_{o1}g}{A_1 \sqrt{2gH_{1N}}}, a_2 = -\frac{A_{o2}g}{A_2 \sqrt{2gH_{2N}}}, \quad (1)$$

$$b_1 = \frac{K_p}{A_1}, b_2 = \frac{A_{o1}g}{A_2 \sqrt{2gH_{1N}}}. \quad (2)$$

Equation 10. Coefficients in linear differential equations

Based on Equation 10 and values from Table 1, linear continuous-time models in the form of transfer functions for the first and second reservoirs are easily determined by applying the Laplace transform to Equation 8.

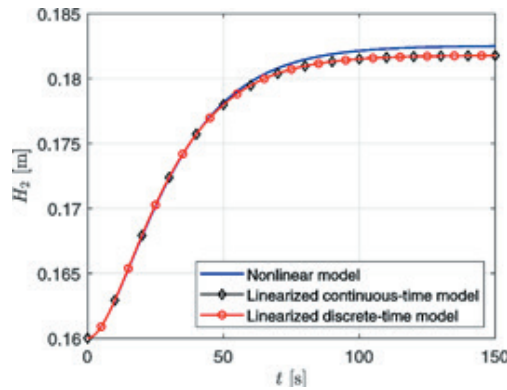
For determining the linear discrete-time model, a sampling period of $T = 0.01$ s is adopted. The Zero Order Hold method is chosen as the discretization method.

Operating conditions and mathematical model representations (both continuous-time and discrete-time) for Tanks 1 and 2 are provided in Table 2.

Figure 2 represents an open-loop plant output comparison of nonlinear with linearized continuous-time and discrete-time models when the deviation from a nominal pump voltage equals $v_p = 0.4$ V. As can be seen from Figure 2, the linear continuous-time and linear discrete-time models perfectly match, which allows for easier design of the controllers in the continuation of the work.

**Table 2.** Nominal operating conditions with dynamic models

	Tank 1	Tank 2
Nominal water level	$H_{IN}=0.16$ m	$H_{2N}=0.16$ m
Nominal pump voltage	$V_{pN}=5.8775$ V	
Linear continuous-time model	$G_1(s) = \frac{0.0544}{15.73s+1}$	$G_2(s) = \frac{1}{15.73s+1}$
Linear discrete-time model	$G_1(z) = \frac{3.4594 \cdot 10^{-5}}{z - 0.9994}$	$G_2(z) = \frac{6.3540 \cdot 10^{-4}}{z - 0.9994}$

**Figure 2.** Continuous-time and discrete-time linear model accuracy against nonlinear plant behavior

4. CLASSICAL PI AND PS CONTROLLERS

Controller input is error signal $e(t)$,

$$e(t) = h_{2d}(t) - h_2(t),$$

Equation 11. Definition of error

and output is control signal $u(t)$ expressed with

$$u(t) = K_p \left(e(t) + \frac{1}{T_I} \int_0^t e(\tau) d\tau \right) = K_p e(t) + K_I \int_0^t e(\tau) d\tau.$$

Equation 12. PI control algorithm

PI controller s -transfer function equals

$$G_{PI}(s) = K_p \left(1 + \frac{1}{T_I s} \right).$$

Equation 13. PI controller s -transfer function

The proportional gain K_p and integral time constant T_I define the key parameters of the controller. Although the coefficients of PID-type controllers are adjustable through various tuning approaches, this study employs the established Ziegler-Nichols method to enable direct comparison with the metaheuristic optimization technique. These same controller parameters K_p and $K_I \equiv K_s$ can be applied in the discrete implementation of the zero-order proportional-sum controller, whose difference equation is described by:

$$u[k] = K_p e[k] + K_s T \sum_{j=0}^{k-1} e[j].$$

Equation 14. Zero-order proportional-sum controller

By eliminating the sum from the previous equation, the final difference equation of the aforementioned PS controller is obtained,

$$u[k+1] = u[k] + K_p e[k+1] + (K_s T - K_p) e[k].$$

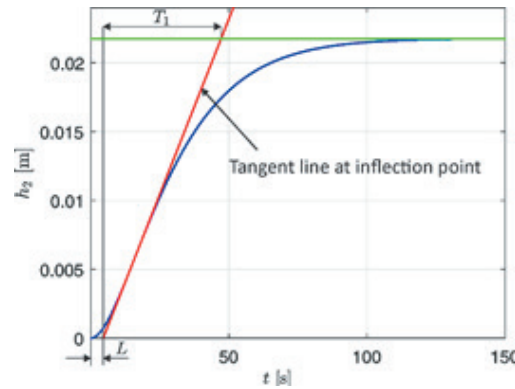
Equation 15. The final form of the difference equation for the PS controller

4.1. ZIEGLER-NICHOLS PI TUNING METHODOLOGY

The investigated plant represents a stable second-order system, making it suitable for the Ziegler-Nichols open-loop tuning method [12]. The procedure begins by obtaining the open-loop step response, which is then analyzed to extract the critical tuning parameters: dead time L , time constant T_I , and process gain K . These parameters are derived as illustrated in Figure 3. Table 3 provides the standard Ziegler-Nichols formulations used to compute the proportional and proportional-integral control parameters.

**Table 3.** PID tuning parameters calculated using the Ziegler-Nichols method

	K_p	T_i	T_d
P	$T_i / (KL)$	-	-
PI	$(0.9T_i) / (KL)$	$3.3L$	-
PID	$(1.2T_i) / (KL)$	$2L$	$0.5L$

**Figure 3.** Tangent method

Lastly, the computed PS controller parameters obtained for the linear model are $K_p=200.8541$, $K_s=16.9559$.

5. THE WHALE OPTIMIZATION ALGORITHM

The whale-inspired optimization method has shown exceptional performance in addressing diverse nonlinear and multimodal challenges. The advantage of this method, and all metaheuristic algorithms in general, is the random distribution mechanism. This distribution helps avoid convergence to local minima. Proposed by Seyedali Mirjalili and Andrew Lewis in [11], WOA mimics the hunting behavior of humpback whales. It is the leader whale's responsibility to locate the fish. The remaining members track directional cues. In every hunt, they all take precisely the same position. They hunt in groups using a three-phase strategy: encircling prey, a bubble-net attack, and an adaptive search. The first phase is to determine the best search agent and update the positions of other agents. Using the distance vectors $\mathbf{D}$ and $\mathbf{X}$ to update the position, the mathematical model of this stage is:

$$\mathbf{A}=2\mathbf{ar}-\mathbf{a}, \mathbf{C}=2\mathbf{r}, \quad (1)$$

$$\mathbf{D}=|\mathbf{CX}'(t)-\mathbf{X}(t)|, \quad (2)$$

$$\mathbf{X}(t+1)=\mathbf{X}'(t)-\mathbf{AD}, \quad (3)$$

Equation 16. Mathematical model of the first phase

where $\mathbf{a}$ is linearly decreased from 2 to 0 and $\mathbf{r}$ is a random vector in $[0, 1]$. $\mathbf{A}$ and $\mathbf{C}$ are coefficient vectors, and t is the current iteration. $\mathbf{X}$ is the position vector, and $\mathbf{X}'$ is the position vector of the best solution so far. The fundamental mathematical models that mimic the second phase are the spiral path (first calculate the distance between the whale and prey using helix movement) and the shrinking encircling mechanism (define the new position of the searching agent using $\mathbf{A}$). Between the original position and the best agent at the moment is the agent's new position. The function of this strategy is

$$\mathbf{X}(t+1) = \begin{cases} \mathbf{X}'(t) - \mathbf{AD} & \text{if } p < 0.5 \\ \mathbf{D}' e^{bl} \cos(2\pi l) + \mathbf{X}'(t) & \text{if } p \geq 0.5 \end{cases}$$

Equation 17. Mathematical model of the second phase

where $\mathbf{D}'$ is the distance between the i -th whale and the prey, l is a random value in $[-1, 1]$, b is a constant for the shape of the logarithmic spiral, and p is a random number in $[0, 1]$. In order to offer adequate connection between the first two phases, the third phase is based on the adaptive variation that relies on the value search vector $\mathbf{A}$.

In this paper, with a population of 35 agents and 25 iterations (determined empirically via trial and error), each agent (whale) encodes a potential solution (optimal parameters K_p and K_s of the controller). Optimization minimizes the sum of squared errors (SSE) as the objective function,



$$SSE = \sum_{i=1}^n e^2 [i].$$

Equation 18. Objective function in the form of the sum of squared errors

Thirty independent runs were executed. A box plot of the objective function values across independent runs is shown in Figure 4 (left). The convergence graph of the best run is shown in Figure 4 (right).

The parameter values generated by the optimization algorithm are $K_P=299.8767$, $K_S=1.9268$.

6. SIMULATION RESULTS

This section presents simulation results demonstrating the plant's response under the control of the various controllers designed earlier. Figure 5 compares the level H_2 control performance between two distinct PS controller implementations. The whale optimization algorithm tunes PS controller gains (K_p and K_s) to outperform the classical Ziegler-Nichols method.

Figure 6 displays the change in control signals for each applied controller. To enhance plant performance, the proposed method generates a superior process input compared to Ziegler-Nichols PS control.

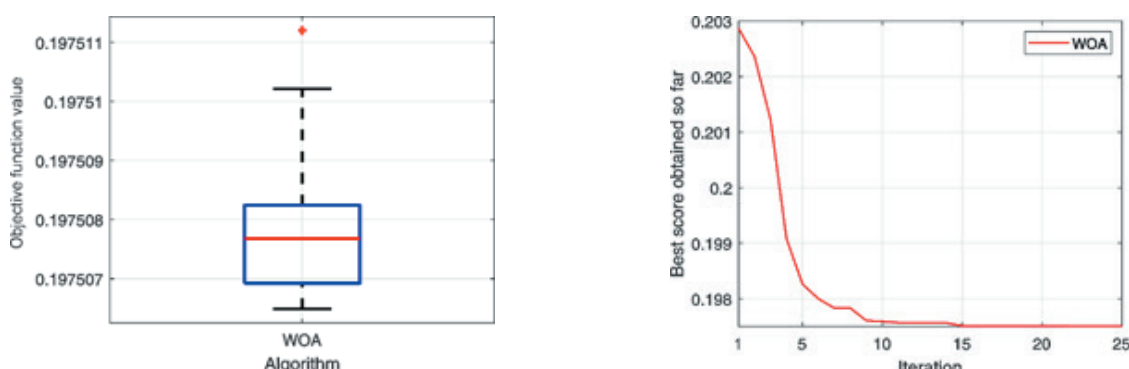


Figure 4. Box plot of 30-run fitness values (left) and convergence graph for the best execution (right)

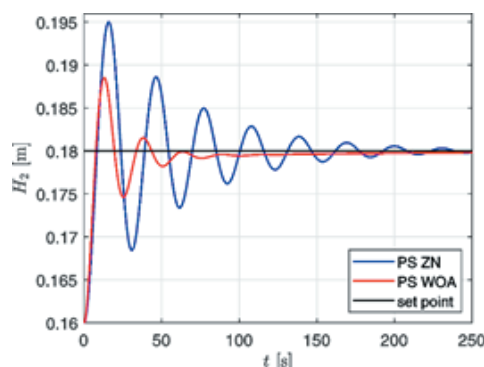


Figure 5. Closed-loop control of H_2 level using multiple PS controllers

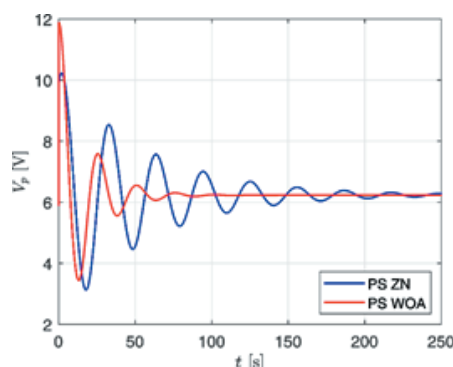


Figure 6. Simulated control signals for proportional-sum controllers

**Table 4.** Numerical values for classical and optimized PS controllers

Controller	SSE	Overshoot [%]	Settling time [s]
PS ZN	0.4839	75	157.03
PS WOA	0.1975	42.5	56.2

To quantitatively demonstrate the optimization's effectiveness, Table 4 presents the calculated sum of squared error values using Equation 18 for the chosen simulation duration. Additionally, in the same table, the system's dynamic behavior was given through both overshoot and settling time ($\pm 5\%$ of the steady-state value). The comparative results in Table 4 indicate that the optimized PS controller outperforms the Ziegler-Nichols PS controller, as evidenced by the shown values.

7. CONCLUSION

In this paper, an analytical nonlinear model of a coupled tank system was obtained. After that, linear continuous-time and discrete-time models were determined. Their comparison was made to show that the linear models accurately describe the nonlinear system around the chosen nominal point. Subsequently, the coupled-tank liquid level process is regulated via a discrete-time proportional-sum control scheme. Gain values were identified employing the Ziegler-Nichols technique alongside the modern whale-inspired optimization approach. To assess the efficacy of their operation, the controllers were systematically compared. Performance comparison of all results was conducted using the minimum sum of squared errors criterion, percentage overshoot, and settling time values. This study highlights the efficacy of metaheuristic optimization in enhancing classical control methods for nonlinear industrial processes. Future studies will focus on the implementation of the obtained results on a real object under laboratory conditions, as well as on testing other optimization algorithms to further improve control performance.

8. ACKNOWLEDGMENTS

This work was financially supported by the Ministry of Education, Science and Technological Development of the Serbian Government, under contract 451-03-137/2025-03/200105, from date 04.02.2025.

REFERENCES

- [1] H. I. Jaafar, S. Y. S. Hussein, N. A. Selamat, M. S. M. Aras, and M. Z. A. Rashid, "Development of PID Controller for Controlling Desired Level of Coupled Tank System," *Internal Journal of Innovative Technology and Exploring Engineering*, vol. 3, no. 9, pp. 32-36, 2014.
- [2] M. Ilyas, S. A. R. Shah, A. Ruf, Y. Khan, and M. Ayaz, "Stabilization of Liquid Level in Tank System Based on Fuzzy Logic Controller," *International Journal of Robotics and Automation*, vol. 11, no. 4, pp. 315-323, 2022.
- [3] S. Ahmad, S. Ali, and R. Tabasha, "The Design and Implementation of a Fuzzy Gain-Scheduled PID Controller for the Festo MPS PA Compact Workstation Liquid Level Control," *Engineering Science and Technology, an International Journal*, vol. 23, no. 2, pp. 307-315, 2020.
- [4] B. S. Sousa, F. V. Silva, and A. M. F. Fileti, "Level Control of Coupled Tank System Based on Neural Network Techniques," *Chemical Product and Process Modeling*, vol. 15, no. 3, p. 20190086, 2020.
- [5] L. A. Torres-Salomao and J. Anzurez-Marin, "Adaptive Neuro-Fuzzy Inference System Control for a Two Tanks Hydraulic System Model," in *2013 IEEE International Autumn Meeting on Power Electronics and Computing*, Morelia, 2013.
- [6] D. Pradeepkannan and S. Sathiyamoorthy, "Control of a Non-Linear Coupled Spherical Tank Process Using GA Tuned PID Controller," in *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, Ramanathapuram, 2014.
- [7] J. Bhokya, M. Vijaya Kumar, J. Ravi Kumar, and A. Seshagiri Rao, "Implementation of PID Controller for Liquid Level System Using mGWO and Integration of IoT application," *Journal of Industrial Information Integration*, vol. 28, p. 100368, 2022.
- [8] R. Jovanović, V. Zarić, M. Vesović, and L. Laban, "Modeling and Control of a Liquid Level System Based on the Takagi-Sugeno Fuzzy Model Using the Whale Optimization Algorithm," in *Proceedings of Papers - 7th International Conference on Electrical, Electronic and Computing Engineering IcETRAN 2020*, Belgrade, 2020.



- [9] N. Budimirovic, E. Prabhu, M. Antonijevic, M. Zivkovic, N. Bacanin, I. Strumberger, and K. Venkatachalam, "COVID-19 Severity Prediction Using Enhanced Whale with Salp Swarm Feature Classification," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1685-1698, 2022.
- [10] S. Golubovic, A. Petrovic, A. Bozovic, M. Antonijevic, M. Zivkovic, and N. Bacanin, "Gold Price Forecast Using Variational Mode Decomposition-Aided Long Short-Term Model Tuned by Modified Whale Optimization Algorithm," in *Data Intelligence and Cognitive Informatics. ICDICI 2023.*, Singapore, 2024.
- [11] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," *Advances in Engineering Software*, vol. 95, pp. 51-67, 2016.
- [12] J. G. Ziegler and N. B. Nichols, "Optimum Settings for Automatic Controllers," *Transactions of the American Society of Mechanical Engineers*, vol. 64, no. 8, pp. 759-765, 1942.



A PLATFORM-AGNOSTIC DEPLOYMENT STRATEGY LEVERAGING REAL-TIME HOST METRICS FOR OPTIMIZED RESOURCE UTILIZATION

Teodor Petrović*,
[0009-0008-7186-2552]

Aleksa Vidaković,
[0009-0005-3527-011X]

Petar Kresoja,
[0009-0008-3343-1540]

Nikola Savanović,
[0000-0001-9670-7374]

Saša Adamović
[0000-0002-2875-685X]

Singidunum University,
Belgrade, Srebia

Abstract:

Many environments for application development face challenges due to inefficient resource utilization and a high degree of vendor lock-in when using cloud services, which limits flexibility and increases user costs. Traditional methods often rely on fixed infrastructures that can lead to host misconfigurations and suboptimal resource usage. In response to these issues, we propose a platform-agnostic application deployment tool that harnesses real-time host metrics to optimize resource allocation and enhance deployment efficiency. The solution features a robust architecture with a client interface, an orchestrator service, a metrics collection service, a dedicated management service, and a proxy service. The platform dynamically configures individual host environments based on specific application requirements by leveraging user-configurable tags and Ansible automation scripts. It further employs host performance metrics, such as CPU and RAM usage and network throughput collected via the metrics collection service, to intelligently select the least utilized hosts for application deployment. Our automated deployment strategy, based on host resource utilization, helps us avoid the traditional issues of having to configure each host manually and not having to be vendor-locked to a specific cloud provider, paving the way for a more flexible and efficient resource utilization.

Keywords:

Platform-agnostic Deployment, Ansible, Metrics-based Resource Allocation, Automated Configuration.

INTRODUCTION

Deploying applications across multiple hosts—whether virtual machines or bare-metal systems—remains a formidable challenge in modern infrastructures. Prior research indicates that inefficient resource utilization is a persistent problem [1]. In addition, containerized environments often suffer from host misconfigurations that adversely affect performance [2]. Traditional deployment methods that rely on manual setup or basic automation scripts lead to further issues such as suboptimal resource usage, insufficient monitoring, and security inconsistencies [3]. Independent studies have shown that these limitations can result in frequent deployment failures [4]. While cloud providers offer a standardized and controlled platform for deploying applications, thus removing the previously mentioned problems, the associated costs and vendor lock-in possibility provide considerable issues with this approach [5].

Correspondence:

Teodor Petrović

e-mail:

tpetrovic@singidunum.ac.rs





To address these challenges, we propose a dynamic, platform-agnostic deployment tool that streamlines application management workflows using available resources. This tool enables users to register their host machines and specify resource and service requirements via an intuitive tagging mechanism. By integrating automated configuration management with continuous, real-time monitoring of host performance, the tool minimizes manual errors and resource wastage while dynamically selecting the optimal host based on current utilization. In doing so, it effectively mitigates issues such as host misconfigurations and vendor lock-in, offering a flexible, scalable, and secure deployment framework.

2. RELATED CONTENT

Previous studies in this field have highlighted significant challenges concerning deploying and managing host resources. Tools such as Kubernetes and Docker Swarm provide robust container orchestration capabilities; however, they typically require complex configuration patterns. Research has demonstrated that these platforms often lack integrated real-time host resource metrics [6] and may not adapt well to fluctuating workloads [7]. Similarly, while Ansible simplifies host configuration through its agentless design, it does not inherently support dynamic, resource-based decision-making, which limits its adaptability in rapidly changing environments. While these platforms have advanced the way containerized applications are managed, their static configuration models sometimes fail to adapt to dynamic resource fluctuations and evolving workload demands in situations where businesses rely on their infrastructure (private or hybrid cloud). This rigidity can lead to inefficient resource allocation and may result in unexpected performance bottlenecks per host machine and the organization as a whole.

Ansible, known for its simplicity and agentless operation, has become a popular choice for host configuration management. Despite its ease of use, Ansible traditionally does not incorporate dynamic, resource-based decision-making, essential for optimizing deployments in highly variable environments. As a result, many deployment scenarios still require manual reconfiguration of the playbook files, which can hinder the system's ability to adapt to evolving infrastructure conditions.

Monitoring systems like Prometheus [8] and Elastic Stack (Kibana) [9] have been instrumental in offering deep insights into application and host performance.

They provide detailed graphs and metrics for visualization and analytics capabilities that allow administrators to track host machines with critical performance statuses over time. However, these systems generally operate in isolation from the deployment process, creating a disconnect between the host monitoring and its automated configuration. To bridge this gap, recent studies have explored integrating real-time performance metrics directly into deployment workflows.

In contrast to these existing solutions, our approach merges real-time monitoring with automated deployment, as demonstrated in recent studies [10]. Additionally, the work by Vankayalapati et al. (2022) [10] supports the concept of predictive scaling through integrated monitoring and orchestration.

Our approach builds upon this research by seamlessly integrating real-time host performance metrics with a custom automated deployment process. By doing so, we significantly enhance the decision-making process regarding host selection and per-host resource utilization, ensuring that deployments are not only automated but also optimized for current infrastructure conditions. This integration addresses the critical gaps left by traditional tools, offering a more adaptive, efficient, secure, and reliable deployment solution. Unlike these existing solutions, which typically rely on static configurations and operate in isolation from real-time monitoring, our approach leverages dynamic metrics to make informed, adaptive deployment decisions, thereby enhancing resource utilization and system responsiveness.

3. METHODOLOGY

Our deployment solution tool is built upon a comprehensive system that integrates several interconnected services to deliver an end-to-end application deployment process. The approach begins with an intuitive user interface that allows system administrators to register host systems (BareMetal or VM machines). During this registration process, each host is assigned descriptive tags that indicate the specific services and configurations required from it, effectively summarizing its intended role and capabilities (Docker, Kubernetes, ingress, egress, storage). This tagging mechanism lays the groundwork for an automated deployment process that tailors to specific hosts, preparing them to meet diverse deployment needs.



Upon host registration, regardless of assigned tags, a metric collection service is installed as part of the default configuration. Simultaneously, each host is automatically enrolled in our DNS system. Using DNS names instead of relying on static IP addresses provides a more robust and flexible way to manage our deployment infrastructure. Using DNS allows us to update host locations or configurations without requiring changes to specific applications or service configurations, improving resilience and reducing management overhead.

Once these foundational configurations are complete, the backend service dynamically generates configuration scripts tailored to each host's designated role or roles, as determined by their tags. These scripts automate the subsequent setup process, significantly reducing manual intervention and the risk of human error or misconfiguration. By leveraging these dynamically generated scripts alongside the standardized metric collection and DNS enrolment, we ensure that each host is optimally and securely configured according to its unique tag profile.

Simultaneously, a metrics collection mechanism continuously gathers real-time performance data from every registered host. Key performance metrics—such as CPU, RAM usage, and network activity—are monitored without interruption. This continuous data flow is critical, as it maintains an accurate, up-to-date view of all host resource utilization across the entire system infrastructure. The collected metrics serve as the foundation for the system's decision-making process.

At the heart of the system lies the management node, which is responsible for orchestrating the deployment process and making calculated decisions based on the latest performance data. When a deployment request is initiated, the management node retrieves the current metrics from the log collection service, filters, and analyzes them to determine the optimal host for a specific application deployment. It employs a decision-making process that calculates the average resource utilization across available hosts, in combination with the application's specific requirements, ultimately selecting the host with the lowest load to ensure efficient deployment and optimal application performance.

The management node not only handles the host selection process but also oversees the entire orchestration process, from initiating host configuration to finalizing application deployment. It acts as a central orchestrator, ensuring that each step of the deployment process from start to finish is executed in the correct sequence and that all system components operate in complete harmony. This coordination is accomplished by pre-defined decision criteria combined with real-time analytics, allowing the system to adapt dynamically to fluctuating resource demands and varying host conditions.

In summary, our methodology integrates automated configuration, continuous real-time monitoring, and intelligent host selection to create a robust deployment process. This integrated approach (see Table 1) not only enhances system reliability and scalability but also directly addresses challenges such as resource inefficiencies and configuration inconsistencies.

Table 1. Service components

Service Name	Function	Technology Stack
Client Interface	Provides an intuitive UI for host registration and configuration	React, JavaScript, HTML/CSS
Orchestrator Service	Manages application deployment orchestration and host selection	Nest.js.
Metrics Collection Service	Gathers real-time performance metrics from hosts	Elasticsearch, Kibana, Logstash, Metricbeat.
Management Service	Automates host configuration and deployment using generated Ansible scripts	Python Flask App.
DNS service	Manages domain name resolution and dynamic DNS records for application access	PowerDNS
Proxy Service	Configures reverse proxy routes for secure application access	Apache HTTP Server



4. IMPLEMENTATION

The system's implementation is designed to automate application deployment by integrating several interconnected components that work in synergy. At its core, the deployment process begins with a default build configuration that defines the default parameters of an application deployment. This build configuration contains the application's git repository URL, branch name, container's port mappings, volume mounts, network assignments, resource limits, restart policies, deployment order, and most importantly, its tags (CPU intensive, RAM intensive, ingress intensive, egress intensive, storage cold, storage hot, etc.). It serves as a deployment blueprint, ensuring that each application is deployed in a consistent and isolated environment while meeting its specific operational requirements.

4.1. DEFAULT CONFIGURATION EXAMPLE

Listing 1 shows an example of a JSON configuration file that defines the deployment parameters for an application. For instance, consider the following build configuration:

As seen in Listing 1, the configuration parameters are explicitly defined to support automated deployment. Based on this configuration, the management node

connects to the specified host system using the SSH key and username, first building the required image for the application. Container images are named using the Git project name, with the tag being the branch name and last commit hash. However, we must note that the Docker command can create previously non-existent volumes if specified with the “-v” option, but there is no command to automatically create new networks. Therefore, if the requested network does not exist, it must be first explicitly created before running the container. Listing 2 represents the Docker run commands automatically generated by the management node based on the configuration provided in Listing 1. Outbound ports are dynamically chosen on the host system, starting from port 10000, based on the host's available free ports. The generated Docker command examples look like this:

In this example, container ports 8080 and 8081 are mapped to host ports 10000 and 10001, respectively. The command sets CPU and memory limits, mounts the specified volume, attaches the container to the primary network (if networks with the requested name do not exist, they will be created), and applies the restart policy.

```
{
  "repository_url": "https://github.com/deploy/web-app",
  "branch": "master",
  "ports": [8080, 8081],
  "tags": ["ingress", "docker"]
  "host": {
    "username": "teodor",
    "domain_name": "nebula-mng.masofinonebula.internal",
    "port": 22,
    "ssh_key_path": "/path/to/key/host_192.168.0.203_Nebula_mng",
    "ip": "192.168.0.203"
  },
  "volumes": [
    { "volumen_name": "SharedData", "container_path": "/home/shared" }
  ],
  "networks": [
    { "name": "sdn-network", "ip": "dhcp" },
    { "name": "internet", "ip": "172.66.0.10" }
  ],
  "resources": { "cpu_limit": "0.5", "memory_limit": "512m" },
  "restart_policy": "always",
  "deployment_order": 1
}
```

Listing 1. Application build config in JSON format



```
# Create the required Docker network
docker network create --internal sdn-network

# Deploy the container based on the configuration in Listing 1
docker run -d
  --name web-app-6 \
  --restart always \
  -p 10000:8080 \
  -p 10001:8081 \
  --cpus="0.5" \
  --memory= "512m" \
  -v SharedData:/home/shared \
  --network sdn-network \
  --network internet \
  --ip 172.66.0.10 \
  web-app-docker:main-b37a3cb
```

Listing 2. Generated Docker run command based on application build config

4.2. DECISION-MAKING ALGORITHM SPECIFICATION

The next step in the deployment process is the selection of the most suitable host for the proposed application. To achieve this, the management node queries and analyses real-time performance metric data from currently active hosts, which fit the application needs, to determine the optimal deployment target. The decision-making process consists of the following steps:

- **Data Collection** - Continuously gathering host performance metrics (e.g., CPU usage, memory consumption, network throughput) from all registered hosts. To account for differences in hardware capacity and application requirements, the raw metrics r_{cpu} , r_{ram} , r_{send} and $r_{receive}$ are normalized using a min-max normalization approach (see Equation 1). Specifically, for each metric, the normalized value is computed as:

$$Metric_{norm} = \frac{r_{metric}}{M_{metric}}$$

Equation 1. Normalization Equation

where M_{metric} is the maximum observed value for that metric.

- **Metric Aggregation** - For each host instance, we calculate a utilization score by combining the normalized metrics. Let CPU_{norm} , RAM_{norm} , $Send_{norm}$, $Receive_{norm}$ denote the normalized values for CPU, RAM, send bytes, and received bytes, respectively. The aggregated score S is then calculated as a weighted average:

$$S = \frac{w_{cpu} \cdot CPU_{norm} + w_{ram} \cdot RAM_{norm} + w_{send} \cdot Send_{norm} + w_{receive} \cdot Receive_{norm}}{w_{cpu} + w_{ram} + w_{send} + w_{receive}}$$

Equation 2. Aggregated Score Equation

where w_{cpu} , w_{ram} , w_{send} and $w_{receive}$ are weights assigned to each metric. In our default configuration, equal weighting is assumed for the base algorithm state. Consequently, when $w_{cpu} = w_{ram} = w_{send} = w_{receive} = 1$, the aggregated score simplifies to:

$$S = \frac{CPU_{norm} + RAM_{norm} + Send_{norm} + Receive_{norm}}{4}$$

Equation 3. Default Aggregated Score Equation

This calculation can be adjusted based on application-specific resource demands (e.g., increasing the weights for CPU and memory if the application is resource-intensive in these areas).

- **Host Ranking** - The aggregated scores of specific hosts that fit the application needs are compared. The host with the lowest overall score is identified, indicating the most readily available host that can handle the new deployment.
- **Decision Execution** - If the build config does not specify a target host, automatically select the host with the lowest utilization score. In the case of a tie or other scenarios (such as deployment order or historical performance data) are defined, we can apply custom decision rules to finalize host selection.

This algorithm ensures that applications are deployed on the most efficient host available at any given moment, thereby optimizing resource utilization and maintaining application stability.



5. EXPERIMENTAL SETUP

The evaluation of the proposed deployment tool was carried out in a controlled environment. The proxy server was hosted on an Ubuntu Linux VPS equipped with four cores and 8 GB of RAM. Each host machine used for deployment had four cores and 8 GB of RAM. The management and log collection machines were provisioned with four cores and 16 GB of RAM, while the backend service ran on a machine with four cores and 8 GB of RAM; all these machines operated on Ubuntu Linux. In addition, all virtual machines—except for the proxy, which is hosted on a VPS—were managed by a hypervisor running VMware Workstation. The hypervisor featured dual Intel® Xeon® E5-2630 v4 processors (2.20 GHz, 10 cores, 10 logical processors each) and 226 GB of RAM. The proxy is connected to the private infrastructure via a WireGuard VPN, ensuring secure and reliable communication between the machines.

6. RESULTS

The deployment tool was evaluated in a controlled environment involving applications with varying resource requirements. We tested the decision-making algorithm on four host instances to assess the efficiency of our metrics-driven host selection process. As the previously defined metrics, we generated aggregated scores based on normalized CPU usage, RAM usage, send bytes, and received bytes. Moreover, our decision-making process is designed in such a way that it can adjust the weights applied to these metrics based on application configuration tags. For instance, if an application specifies high CPU and RAM requirements, the algorithm increases the weighting factors for these metrics, ensuring that hosts with lower resource utilization in these areas are preferred. Similarly, significant ingress and egress network requirements can lead to higher send and received byte weights. Table 2 summarizes the normalized performance metrics and aggregated scores obtained from our experiments on four host instances.

Based on the results in Table 2, the decision-making algorithm selected abyss-mng as the best deployment target because it scored the lowest score (0.25). This outcome demonstrates that our metrics-driven approach effectively reduces resource wastage and enhances resource utilization, all while aligning with application-specific resource demands.

7. CONCLUSION

The developed platform-agnostic deployment tool successfully integrates automated host configuration, metrics-driven resource allocation, and user-driven deployment customization. Based on the experimental results presented in Table 2, our evaluation shows that the system effectively selects the best host for deployment by considering host performance metrics—including CPU, RAM memory, send bytes, and received bytes—and dynamically adjusting to application-specific requirements. In conclusion, we have successfully achieved our objectives by addressing traditional deployment challenges such as inefficient resource utilization and misconfiguration. The system not only minimizes resource wastage but also significantly enhances deployment stability and scalability, proving its effectiveness in real-world scenarios. Future work will focus on integrating machine learning techniques to further refine the adaptive weighting mechanism and on scaling the system to handle a larger number of hosts under varying workload conditions.

Table 2. Normalized Performance Metrics and Aggregated Scores for Host Selection

Host	CPU Usage	Memory Usage	Send Bytes	Received Bytes	Score
madman-mng	0.50	0.52	0.55	0.56	0.53
abyss-mng	0.28	0.31	0.22	0.18	0.25
nexus-mng	0.40	0.38	0.35	0.33	0.37
nebula -mng	0.42	0.37	0.36	0.35	0.38



REFERENCES

- [1] M. Narasimhulu, D. V. Mounika, P. Varshini, A. K. and T. R. K. Rao, "Investigating the Impact of Containerization on the Deployment Process in DevOps," *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pp. 679-685, 2023, <https://doi.org/10.1109/ICECAA58104.2023.10212240>
- [2] M. Abhishek, D. Rao and K. Subrahmanyam, "Framework to Deploy Containers using Kubernetes and CI/CD Pipeline," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 239, 2022, <http://dx.doi.org/10.14569/IJACSA.2022.0130460>
- [3] W. M. C. J. T. Kithulwatta, W. U. Wickramarachchi, K. P. N. Jayasena, B. T. G. S. Kumara and R. M. K. T. Rathnayaka, "Adoption of Docker Containers as an Infrastructure for Deploying Software Applications: A Review," *Advances on Smart and Soft Computing*, p. 247–259, 29 June 2021, https://doi.org/10.1007/978-981-16-5559-3_21
- [4] M. Dînga, L. Giamattei, A. Guerriero, R. Pietrantuono, S. Russo, I. Malavolta, T. Islam, M. Dînga, A. Koziolk, S. Singh, M. Armbruster, J.-M. Gutierrez-Martinez, S. Caro-Alvaro, D. Rodriguez, S. Weber, E. F. Vogelin and F. S. Panojo, "Monitoring tools for DevOps and microservices: A systematic grey literature review," *Journal of Systems and Software*, vol. 208, no. 0164-1212, p. 111906, 16 December 2024, <https://doi.org/10.1016/j.jss.2023.111906>
- [5] D. Mo, R. Cordingley, D. Chinn and W. Lloyd, "Addressing Serverless Computing Vendor Lock-In through Cloud Service Abstraction," *2023 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 193-199, 2023, <https://doi.org/10.1109/CloudCom59040.2023.00040>
- [6] "Cloud resource orchestration in the multi-cloud landscape: a systematic review of existing frameworks," *Journal of Cloud Computing*, vol. 9, no. 49, 2020, <https://doi.org/10.1186/s13677-020-00194-7>
- [7] S. Son and Y. Kwon, "Performance of ELK stack and commercial system in security log analysis," *2017 IEEE 13th Malaysia International Conference on Communications (MICC)*, pp. 1-6, 2017, <http://dx.doi.org/10.1109/MICC.2017.8311756>
- [8] Y. Liu, Z. Yu, Q. Wang, H. Mei, G. Song and H. Li, "Research on cloud-native monitoring system based on Prometheus," *Fourth International Conference on Sensors and Information Technology (ICSI 2024)*, vol. 13107, p. 131071B, 2024, <https://doi.org/10.1117/12.3029320>
- [9] A. S. Shaji and M. M. George, "Elastic Stack: A Comprehensive Overview," *2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 1-5, 27 July 2024, <https://doi.org/10.1109/RAICS61201.2024.10690099>
- [10] R. K. Vankayalapati, A. Edward and Z. Yasmeen, "Composable Infrastructure: Towards Dynamic Resource Allocation in Multi-Cloud Environments," *Universal Journal of Computer Sciences and Communications*, vol. 1, 2022, <https://doi.org/10.31586/ujcsc.2022.1222>



LEVERAGING LLMs FOR AUTOMATIC FORUM SCRAPER GENERATION

Miloš Pavković^{1*},
[0000-0001-7776-6045]

Jelica Protić²,
[0000-0003-0846-0290]

Petar Kresoja¹
[0009-0008-3343-1540]

¹Singidunum University,
Belgrade, Serbia

²School of Electrical Engineering,
University of Belgrade,
Belgrade, Serbia

Abstract:

Web forums contain valuable user-generated content (UGC), but crawling them presents a challenging task due to the differences between forum technologies and structures. This paper proposes a general approach that uses Large Language Models (LLMs) to automatically detect the forum technology (e.g., phpBB, vBulletin, SMF, Discuz!) and generate a web scraper for that forum's layout, structure, and pagination. LLM first identifies the platform of a given forum by analysing its HTML patterns after it generates code to efficiently collect available posts and threads that are publicly available and don't require user registration.

Several state-of-the-art LLMs are evaluated (GPT-4, Claude 2, and Mistral 7B) for this task, comparing their speed, accuracy, and reliability in generating functional scraping code. A proof-of-concept functionality was demonstrated on a chosen phpBB forum technology by crawling its content with LLM-generated Python code.

Experimental results show that the LLM-generated scrapers can successfully retrieve forum posts with high accuracy, matching manually coded crawlers while adapting automatically to different forum structures. The findings suggest that LLMs can significantly improve forum data collection, avoiding manual per-site adjustments and reducing duplicate content in incremental crawls.

Keywords:

Large Language Models, Web Scraping Automation, Template Detection, Data Retrieval.

INTRODUCTION

Online discussion forums are a valuable source of user-generated content, containing discussions, Q&A, reviews, and community knowledge [1]. Retrieving this content has applications in domains like social media analysis, customer feedback evaluation, and knowledge extraction. However, forums present unique challenges for web crawlers: content is spread across pages, threads and posts, and each forum technology (phpBB, vBulletin, SMF, Discuz!, etc.) has its own HTML layout and navigation structure. New posts continually shift older posts to new pages, complicating incremental crawling and often leading to duplicate data retrieval if not handled properly. Traditional forum crawlers either rely on manually written code for each forum technology or attempt a one-size-fits-all strategy that may not capture details related to that specific website.

Correspondence:

Miloš Pavković

e-mail:

mpavkovic@singidunum.ac.rs



SInFo [1], a recent structure-driven forum crawler, highlights these issues: it targets the latest content by leveraging forum-specific URL patterns and pagination routes while remaining generic across platforms. SInFo achieved an average of 92.6% new content per recrawl cycle, demonstrating the importance of understanding forum structure to avoid redundant downloads. Despite such advances, implementing a new scraper for each forum or generalizing across all forums still demands significant human effort in analysing HTML and building site specific parsing rules.

In this paper, a new LLM-driven approach for forum crawling is proposed. The key idea is to combine the LLMs understanding of text and code generation with the structured nature of forums. The system first detects the forum technology on a target website by examining distinctive technology features (such as footer text, URL signatures, and HTML layout structure). Once identified, prompts are sent to the LLM to generate a scraping script specific to that platform's structure (e.g., how threads, pages and posts are organized). This two-step approach ensures that details of each forum technology are properly captured and parsed. For example, if the forum technology is recognized as phpBB, the LLM can use knowledge of phpBB's page numbering system and thread HTML structure to produce an accurate crawler. If instead it's an SMF forum, the LLM would know to look for *index.php?topic=* patterns and the corresponding navigation scheme that are unique for this type of technology. This approach is evaluated using three different popular LLMs – GPT-4 by OpenAI [2], Claude 2 by Anthropic [3], and Mistral 7B [4] (a smaller open-source model) – to compare their performance in code generation for this task.

This work, (i) introduces a standardized approach for forum crawling that automatically adapts to different forum technology software using LLMs, (ii) presents a comparison of multiple popular LLMs (proprietary and open source) in generating web forum scraping code in terms of speed, accuracy, and reliability, and (iii) demonstrates through an experimental setup that an LLM-generated scraper can successfully crawl a real forum, matching the efficiency of manually written crawlers. The proposed approach can target only publicly viewable forums, focusing on open content – user logging is explicitly excluded so as bypassing the CAPTCHAs and other restrictions.

2. RELATED WORK

The recent study of focused web crawling advancements has been made by integrating semantic analysis and optimization algorithms to enhance performance. For instance, Liu et al. introduced a focused crawler that combines a semantic disambiguation graph with a semantic vector space model to improve the retrieval of topic-relevant web pages [5]. Similarly, Huang et al. proposed a crawler that constructs a semantic graph to eliminate ambiguous terms and employs a genetic algorithm to optimize weighting factors, resulting in improved acquisition rates and relevance [6]. While these approaches demonstrate the potential of semantic understanding and intelligent learning in focused crawling, they may not fully address the challenges of dynamically detecting and adapting to various forum structures and technologies without extensive manual configuration.

In early research on web forum crawling the challenges were complex navigation and duplicate content. FoCUS (Forum Crawler Under Supervision), which used machine learning, identified forum-specific URL patterns and page templates [7] to handle complex navigation. FoCUS learned regular expressions for forum thread URLs and navigational links by training on annotated forums, enabling it to crawl forums at scale with minimal updates and rules. However, it required supervised training and did not generalize to unseen forum layouts without additional examples [7]. On the other hand, structure-driven crawling methods used the predictable layout of forums: for example, the work of Pavković and Protić on SInFo (Structure-Driven Incremental Forum crawler) proposed a generic two-phase strategy [1]. In SInFo, the crawler first separates index pages (listing threads) from content pages (showing posts), then uses the forum's navigational structure to find the page containing the newest content. By observing the URL format for each forum technology, SInFo could target directly the latest posts, avoiding re-fetching pages seen in previous crawls.

The LLMs ability to understand language, text and code has started the interest in using them to automate web scraping tasks. The core advantage of LLMs is their ability to understand context and generalize on it, which helps in interpreting various webpage structures. LLM-powered scrapers can interpret and understand complex website structures, making them more effective than traditional scraping tools [8]. Unlike traditional scrapers that can easily break when the site's layout changes since they rely on delicate CSS selectors or XPath, LLM-based scrapers leverage natural language understanding to adapt to changes.



Several practical systems have been developed that integrate LLMs into the scraping pipeline. ScrapeGhost by Turk is an experimental library where the user provides a target URL and a desired output schema; the library then prompts GPT-4 to extract the target data [9]. It removes the need to manually write parsing code for each site. Similarly, FireCrawl [10] is an open-source tool that crawls a website and returns content in a clean, structured format (like Markdown or JSON), suitable for feeding into LLM applications. In a more general sense, frameworks like LangChain [11] started including web browsing and scraping capabilities through LLMs, where the model can be instructed to use a browser tool to navigate pages and then parse them with its internal reasoning [12]. These agentic approaches like AutoGPT or BabyAGI can crawl web pages by iteratively deciding which link to follow next and when to stop, using the LLM's output as the controller. While flexible, they often experience high token usage and latency, and need careful prompting to stay focused to avoid the agent getting lost on irrelevant links.

3. METHODOLOGY

The approach proposed in this work consists of two main stages: (A) Forum Technology Detection and (B) Scraper Code Generation, followed by an execution and evaluation phase. In the following chapters, each stage will be described in detail, including how LLMs are utilized and tuned for the task.

3.1. FORUM TECHNOLOGY DETECTION

The first step is to determine which forum technology a target website is running. This is a crucial step because it impacts how page navigation works (for example, phpBB vs. vBulletin have different URL schemes for threads and pages) [1]. The input to the LLM can be a snippet of HTML (such as the forum's front page or a thread page) or extracted textual parts from the page. In many cases, forums explicitly state their platform in the footer – e.g., “Powered by phpBB” or “Powered by SMF” as shown in Figure 1.

If such a signature exists, a simple automatic keyword check is sufficient. In the absence of explicit text, there are other indicative signs that can be used:

- **URL patterns:** phpBB URLs often include *viewtopic.php* or *viewforum.php* with parameters *f* (which stands for forum id) and *t* (for topic id). vBulletin 3/4 uses *showthread.php* and *forumdisplay.php*, whereas vBulletin 5 and some others use SEO-friendly paths but still might contain vbulletin in HTML comments or JavaScript code. SMF (Simple Machines Forum) typically uses *index.php?topic=* for threads and *board=* for sections, and Discuz! (a popular and widespread Chinese forum technology) uses URLs like *forum-<id>-1.html* and *thread-<tid>-<page>-1.html*. An LLM can be provided with a URL or HTML code snippet and prompted with the question: “Identify which forum software this site likely uses.” Thanks to patterns seen during training, models like GPT-4 or Claude 2 can accurately classify the technology analysing forum software specific features. For example, experiments that were done in this work were in the form of feeding GPT-4 with the HTML head and part of the body of a phpBB forum page; the model correctly responded that the site was phpBB, noting the presence of phpBB specific features and typical structure.
- **HTML structure and keywords:** Each forum software has a default layout and often unique element IDs or class names. phpBB's HTML might contain references to classes like *postbody* or form fields with names like *sid* (session id) specific to phpBB. SMF pages show a distinctive table-like structure layout with user info sidebar and usually use labels like “Logged” under posts. Discuz pages might contain Chinese locale strings or certain script names. For the experiments in this work, a small prompt for the LLM known features of common forum technologies is compiled (e.g., “If you recognize 'post.php?action=post' and 'SMF' in the HTML,

Powered by SMF 2.0.17 | SMF © 2006-2011, Simple Machines LLC

XHTML | RSS | WAP2

Powered by phpBB® Forum Software © phpBB Limited

Privacy | Terms

Figure 1. An example of the explicitly stated platform of forum technology



it's SMF; if you recognize 'phpbb/templates', it's phpBB"). The idea of these types of prompts is to act as a few-shot guide. Upon providing the page HTML, the LLM outputs the guessed forum type with high confidence.

In cases where the forum is highly customized or built from scratch, the LLM might output "unknown/custom platform." This itself is useful feedback and it might also indicate that automated generation could be less reliable due to unrecognized structure. For such cases, the proposed approach defaults to a generic strategy (like SInFo [1] approach) or require human confirmation. In shown experiments, detection was straightforward on known forums – all three LLMs correctly identified the technology when clear signatures and software indicators were present. GPT-4 and Claude 2 even correctly identified the version of the forum technology analysing HTML structure indicators and noting "this is a phpBB 3.x forum", whereas the smaller Mistral model sometimes needed the explicit "Powered by" text to be sure.

3.2. SCRAPER CODE GENERATION

Once the forum type is identified, the prompt for the LLM to generate a scraper code is adapted to that specific forum technology. For each target technology, a distinct prompt template is designed, embedding general instructions on how to scrape forums. An example prompt (simplified) for phpBB is shown on Listing 1.

GPT-4 does not even requires a detailed template per platform – it often knows from its trainings the default behaviours. For instance, GPT-4 generated code that searched for a "Next" button or a `&start=` parameter for phpBB after seeing the page HTML code. However, to ensure reliability, the hints are provided in the prompt. For phpBB: "Note: phpBB thread pages use a `start=<n>` parameter for pagination. The first page might have `start=0` (implicit) and subsequent pages `start=15, 30, ...` etc., typically 15 or 20 posts per page. Use this knowledge to iterate through pages until no more posts can be found." For SMF: "SMF thread URLs contain `topic=<id>.<offset>`. E.g., `topic=123.0` for the first page, then `topic=123.15` for the next if 15 posts per page.

Use the presence of a "next" link or increment the offset accordingly." By giving such technology-specific guidance, even a smaller model with fewer parameters like Mistral can follow the correct approach.

The LLM outputs code in Python, but any language could be requested, where Python is chosen for ease of readability and execution. The generated code typically includes: (1) sending an HTTP GET request to the initial URL, (2) parsing the HTML to extract the posts (using BeautifulSoup [13] or similar), (3) finding the URL or parameter for the next page, and repeating this process until no next page is found, and (4) storing or printing the extracted data (e.g., as JSON or CSV). The model is also instructed to include basic error handling (e.g., check response status, limited number of retries and break if a page request fails) to improve robustness.

When prompted for phpBB, GPT-4 correctly used the phpBB-specific classes (*postbody*, *author*, *content*) which it likely picked up from context or training knowledge. It also identified that the "Next" page link is a literal "Next" text anchor in phpBB default template. It is worth noting that GPT-4 was not prompted with specific class names – it reasoned them by itself, demonstrating the model's internal knowledge of phpBB HTML structure.

For models like Claude 2, the approach was similar. Claude tended to be very redundant with comments and sometimes over-engineered the solution (e.g., writing separate functions to parse a page). Compared to GPT-4, Claude also produced correct logic and even handled edge cases correctly like when a page has no "Next" link. On the other hand, Mistral 7B model, being a much smaller model, had experienced greater difficulty. Its first attempt at generating code often missed details (for example, it might not find the correct post container div, or it may stop after one page due to not recognizing the pagination element). By refining the prompt or providing an example HTML snippet to Mistral, it could correct itself to succeed in simple cases, but it was less reliable as out-of-the-box scraper code generation tool. This highlights a trade-off between large proprietary models and smaller open-source ones, which will be evaluated in the results.

```
"You are an expert web scraper. Write a Python script using requests and BeautifulSoup to
scrape all posts from a given phpBB forum thread URL. The script should handle pagination
by finding the 'next page' link or appropriate page parameters, and collect the author,
timestamp, and text of each post. Assume the forum is public (no login needed)."
```

Listing 1. An example of an LLM code generation prompt



3.3. EXECUTION AND DATA COLLECTION

After obtaining the code from the LLM, the next step is to run the scraper and collect the data. For this work and experiments, the generated script was manually executed to verify the correctness and to collect basic metrics. The focus of the evaluation was: Does the script successfully retrieve all posts from the forum (or thread)? And if so, how efficient is it (in terms of requests made or duplicates avoided)?

The advantage of having the forum exact type technology is the potential to optimize the crawling since the traversing structure can be known in advance. For instance, knowing the URL pattern, the LLM could decide to construct the URL for the last page of a thread (using information about total posts). SInFo did this by calculating the page index for the latest content [1]. Experiment with prompting GPT-4 to incorporate such logic in one case, was in the form of instruction “First fetch the thread’s last page to get the latest posts, then retrieve earlier pages if needed and if they exist.” GPT-4 responded with a complex approach which complicated the code. For simplicity and reliability, the main implementations in experiments remain regular forward pagination (page 1 → page 2 → ...). This returns all posts without needing to guess positions and is easier to validate for completeness.

It is important to note that while the LLM could in theory parse the HTML itself (without generating code) – as done in some end-to-end LLM scraping demos [12], having a concrete script offers persistence. The script can be reused to crawl the forum regularly, or shared, without requiring an LLM each time. Here, the LLM is observed as a “crawler generator”. The main workload of the data extraction after that can be done by the code,

which is efficient in execution, not the LLM. This addresses a common concern that LLM as a scraper might be slow or expensive if used for every page or frequently on a very large forum. Here, the LLM’s cost is one time (per site) to produce a scraper code, comparable to a human developer writing it, but much faster.

3.4. LLMS COMPARED

Three different LLMs to the above tasks were observed on their differences: GPT-4, Claude 2, and Mistral 7B. Table 1 provides a summary comparison. GPT-4 (through OpenAI API, 8k context version) is the largest and generally most capable model that was tested in this work. Claude 2 (Anthropic model, ~100k context) is also a top-tier model with the notable advantage of a much larger context window. This in principle allows feeding it with a whole forum HTML page (tens of thousands of tokens) in one request. Mistral 7B is an open-source model released in late 2023. While not specifically a code model, in this work, the instruct variant was used with some success.

The prompting methodology was kept as consistent as possible. The same high-level prompts were used and only the details were adjusted necessary for each model (e.g., simplifying instructions for Mistral due to its smaller capacity). For GPT-4 and Claude, the temperature was set to 0 to minimize randomness, ensuring the outputted code is deterministic and focused. Mistral, being less deterministic even at low temperatures, sometimes generated different codes. The Mistral model ran a few times to take the best outcome for fairness.

Table 1. Comparison of LLMs for forum scraper generation

LLM Model	Size/Type	Speed (gen. tokens)	Code Accuracy	Context Window	Notes
GPT-4 (OpenAI)	~180B, Proprietary	~2-3 tokens/sec (API)	Very High – correct on first try for all tests< robust logic	~8K tokens (std. 32K variant available)	Best reasoning abilities; knows common forum patterns internally; slower and costlier.
Cloude 2 (Anthropic)	~120B, Proprietary	~5-7 tokens/sec (API)	Very High – correct in first try (small fix needed for one case)	100K tokens	Very large input capacity (good for long HTML); slightly more natural code comments; fast generation.
Mistral 7B (Open)	7B, Open-source model	~20+ tokens/sec (local GPU)	Moderate – needed iterative prompting; prone to minor mistakes	~4K tokens (instruct variant)	Runs locally (no API needs); much weaker out-of-the box knowledge; can succeed with guided prompts.



4. EXPERIMENTS

Experiments were conducted to evaluate the accuracy and reliability of the LLM-generated scrapers, and to compare LLM performance in terms of speed and quality for this task. Representative forums were selected for each of the four target technologies (phpBB, vBulletin, SMF, Discuz). The detailed report was only on phpBB and SMF cases; the other two technologies were done in a more limited fashion due to time and scope of this paper, but the results were similar to phpBB and vBulletin.

For testing phpBB, the “Everything Search Engine” forum (voidtools.com/forum) was used. This forum statistics (as of March 2025) indicate over 21,000 posts in the support section alone, making it a robust test for scraping. For SMF, the “Fractal Softworks” forum was used (fractalsoftworks.com/forum), which is an SMF forum with ~109k posts in its general discussion board. Both forums are open access, no login is required to read. The approach was also verified on a smaller vBulletin forum and a Discuz! forum board with an English interface.

Testing pipeline was: feed an HTML snippet to LLM for detection, then prompt for code generation, then execute the code. The following were measured: (a) Was the platform correctly identified? (b) Did the code run without errors? (c) Did it successfully retrieve the expected number of posts? (d) How many HTTP requests did it use, and did it avoid unnecessary pages? And (e) time taken for the LLM to generate the code and for the code to run.

5. RESULTS

LLM Detection and code generation: All tested LLMs correctly identified the forum software in all cases where a clear signature was present (phpBB and SMF tests). For vBulletin, the HTML snippet that was provided was less explicit (the forum had removed the “Powered by” footer), but GPT-4 still reasoned it was vBulletin (likely from a meta tag and form field names), whereas Mistral misclassified it as “maybe phpBB or custom.” This indicates that larger models have an advantage in hard to detect recognition tasks. Once the forum technology was known, GPT-4 and Claude 2 generated working code on the first try for both phpBB and SMF forums. Table 2 shows the scraping results. For the phpBB forum, the GPT-4 generated script successfully scraped all 4,008 threads and 21,239 posts from the “Support” board, by iterating through 40 pages of

thread listings and then crawling each thread’s pages. The entire crawl was completed in about 15 minutes. Claude 2 script was equally successful on phpBB. On the SMF forum, both models also managed to scrape the target board (General Discussion with ~7,906 topics and 109,942 posts). One minor issue arose on Claude for the SMF where initially generated code didn’t construct the next page URL correctly (SMF requires adding an offset like .15 to the topic parameter). In both forums, Mistral 7B eventually produced a functional scraper but only after iterative prompting and guidance.

Speed and Efficiency: In terms of generation speed, GPT-4 was slower (it took ~30 seconds to output ~60 lines of code). Claude 2 was faster, producing similar length code in ~10 seconds. Mistral (running on a local machine with 1xA100 GPU in this test) was extremely fast in generating code – on the order of 20 tokens per second – but since it needed multiple tries, the total time to get the correct code was a couple of minutes. The scraping runtime for each script was comparable since they all used Python requests: the differences came down to how many requests were made. All scrapers successfully followed pagination and did not get dead locked in loops or miss pages. The number of page fetches roughly matched the number of pages in each forum section plus each thread. This is quite efficient given the scale, and importantly, it is the minimal required to get all posts.

Accuracy of Data: The scraped data were checked for errors. For instance, the first and last posts of certain threads were compared to the live website. In both cases, they matched exactly, including formatting. Both GPT-4 and Claude scrapers achieved 100% post recall on the tested forums. When finally worked Mistral code also got all posts from a couple of test threads, but it wasn’t run on the full forum due to lower confidence in its generalization.

LLM Comparison: Table 1 summarizes the comparison among GPT-4, Claude 2, and Mistral 7B. GPT-4 demonstrated the highest reliability, consistently producing correct and well-structured code. It also tended to handle unexpected site anomalies better. Claude 2 advantages were its speed and extremely large context window. A big quality gap was not noticed between the models, both were excellent, with Claude making some minor mistakes. Mistral 7B clearly fell behind in understanding and needed more explicit instructions. Its main benefit is being open-source, so it can be self-hosted and fast.

**Table 2.** Results of LLM-generated scrapers on example forums

Forum (Sowftware)	LLM Usedfor Code	Post/Threads Retrieved	Page Fetched	Success?	Comments
Voidtools "Everything" forum (phpBB)	GPT-4 (Claude 2)	21,239 posts in 4,008 threads (100% of forum section)	~8,000 HTTP requests (40 thread-list pages + thread pages)	Yes	Completed in ~15 min. Claude 2 got same results. No duplicates: matched forum statistics.
Fractal Softworks forum (SMF)	GPT-4 (Claude 2)	109,942 posts in 7,906 topics (full board)	~7,906 thread pages + navigation page (~8,200 total)	Yes	GPT-4 code handles SMF pagination correctly. Claude needed one minor prompt adjustment then succeeded.
Tech Discussion forum (vBulletin 4)	GPT-4	~5,000 posts in 200 threads (sampled)	200 thread pages + index pages	Yes	Partial crawl (forum is large). LLM detection identified vBulletin. Scraper worked on first attempt.
Sample Discuzi forum (Discuz)	GPT-4	~500 posts in 50 threads (test sample)	50 thread pages + index pages	Yes	Discuz specific page pattern followed (thread *.html). Confirmed post count manually.
Overall		All public posts scraped with high accuracy	Efficiency near optimal		LLM-generated scrapers achieved complete data retrieval for each tested forum.

Table 2 highlights the end-to-end effectiveness on test forums. Overall, GPT-4 and Claude 2 are both excellent choices for implementing this type of scraping. Mistral 7B, while not as out of the box solution, is a promising sign that even lightweight models, which will only improve in coming years, could handle such tasks, especially if specialized code focused versions are used.

6. CONCLUSION

In this work, an LLM-based crawling approach is introduced that automates the detection and scraping of forums across technologies like phpBB, SMF, vBulletin, and Discuz, reducing manual effort. The results indicate that an LLM-driven approach to forum crawling is not only feasible, but also significantly influential. Unlike traditional crawlers, this method uses LLMs to generate customized scraping code based on forum structure. Experiments show that GPT-4 and Claude 2 produce accurate, efficient scrapers, while even smaller models like Mistral 7B can succeed with guidance. This lowers the entry barrier for web data collection and demonstrates the real potential of AI-assisted code generation. The development time for a new scraper is reduced from potentially days, if done manually, to minutes with proper LLM prompts. Moreover, the ability of LLMs to generalize well means that even if a forum theme or layout changes, the scraper code will still be correctly generated. The approach is extensible to other semi-structured web domains and opens paths for self-repairing, scalable scraping systems.

In conclusion, this experiment demonstrates that LLMs can serve as powerful allies in web forum crawling, automating what used to be time consuming engineering work.

REFERENCES

- [1] M. Pavkovic and J. Protic, "SInFo – Structure-Driven Incremental Forum Crawler That Optimizes User-Generated Content Retrieval," *IEEE Access*, vol. 7, p. 126941–126961, 2019, doi:10.1109/ACCESS.2019.2939872.
- [2] OpenAI, "GPT-4o," [Online]. Available: <https://openai.com/index/gpt-4-research/>. [Accessed 5 March 2025].
- [3] Anthropic, "Claude 2," [Online]. Available: <https://www.anthropic.com/news/claude-2>. [Accessed 5 March 2025].
- [4] M. AI, "Mistral 7B," [Online]. Available: <https://mistral.ai/news/announcing-mistral-7b>. [Accessed March 2025].
- [5] W. Liu, Y. He, J. Wu, Y. Du, X. Liu, T. Xi, Z. Gan, P. Jiang and X. Huang, "A focused crawler based on semantic disambiguation vector space model," *Complex & Intelligent Systems*, vol. 9, no. 1, p. 345–366, 2023, doi:10.1007/s40747-022-00707-8.
- [6] W. Huang, X. Li, X. Zhou, D. Qi, J. Xi, W. Liu and F. Zhao, "A Semantic and Optimized Focused Crawler Based on Semantic Graph and Genetic Algorithm," *Symmetry*, vol. 16, no. 11, p. 1439, 2024, doi:10.3390/sym16111439.



- [7] J. Jiang, X. Song, N. Yu and C.-Y. Lin, "FoCUS: Learning to Crawl Web Forums," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, p. 1293–1306, 2013, doi:10.1109/TKDE.2012.56.
- [8] reddit, "LLM Powered Web Scrapers Experience," [Online]. Available: https://www.reddit.com/r/LocalLLaMA/comments/1d5q6o7/llm_powered_web_scrapers_experience/. [Accessed 10 March 2025].
- [9] "Scrapeghost," [Online]. Available: <https://simon-willison.net/2023/Mar/26/scrapeghost/>. [Accessed 25 March 2025].
- [10] Firecrawl, "Turn websites into LLM-ready data," [Online]. Available: <https://www.firecrawl.dev/>. [Accessed 25 March 2025].
- [11] LangChain, [Online]. Available: <https://www.lang-chain.com/>. [Accessed 25 March 2025].
- [12] N. Corcuera Platas, "Enhancing Web Scraping With Large Language Models: A Modern Approach," 26 April 2024. [Online]. Available: <https://medium.com/@ignacio.cplatas/enhancing-web-scraping-with-large-language-models-a-modern-approach-6216d5bba8d5#>. [Accessed 25 March 2025].
- [13] "beautifulsoup4," [Online]. Available: <https://pypi.org/project/beautifulsoup4/>. [Accessed 25 March 2025].



INFORMATION TECHNOLOGY SESSION





CASE STUDY OF PERFORMANCE ON THE KVM HYPERVISOR-BASED VIRTUALIZATION RELATED TO NATIVE HOST

Borislav Đorđević*,
[0000-0002-6145-4490]

Kristina Janjić,
[0009-0005-0750-6105]

Nenad Kraljević
[0009-0008-7684-5444]

School of Electrical and Computer
Engineering,
Belgrade, Serbia

Abstract:

This study examines how file system performance differs between a native operating system and a KVM hypervisor-based virtualized environment. The research uses CentOS 9 as both the native/guest OS and employs Filebench for benchmarking purposes. Tests were conducted on both the native OS and within KVM virtual environments configured including one, two, three, and four virtual machines. The study establishes a mathematical model to compare performance between the native and virtual environments. According to the model, the native operating system exhibits significantly faster performance compared to its virtualized counterpart. Empirical results of the paper confirm this prediction, demonstrating a noticeable drop in file system performance.

Keywords:

KVM, Virtualization, Native OS, Linux, Centos 9, Virtual Machine.

INTRODUCTION

Virtualization technologies represent pivotal advancements in Information Technology and Cloud Computing. Hypervisor-based virtualization, the predominant form, enables running multiple full operating systems simultaneously on a single physical machine, facilitated through virtual machines (VMs). There are several types of hypervisor virtualization, including full hardware virtualization, paravirtualization, and operating system-level virtualization. Among these, full hardware virtualization is the most widely adopted. Virtualization offers numerous benefits over traditional single-operating-system architectures. One of the primary advantages is enhanced CPU utilization on physical servers. By enabling the simultaneous execution of multiple operating systems, virtualization optimizes hardware utilization and reduces energy consumption [1]. This capability is fundamental in modern computing environments, facilitating efficient resource allocation and scalability in IT infrastructures and cloud services.

Correspondence:

Borislav Đorđević

e-mail:

borislav.djordjevic@viser.edu.rs





Virtualization indeed brings numerous advantages, but one significant drawback is the noticeable performance decrease for virtual operating systems compared to native environments. In a traditional architecture, a native operating system operates straight on the underlying hardware, achieving optimal performance. However, when the same operating system runs within a virtualized environment, it accesses hardware resources through the hypervisor and the host operating system, leading to a significant performance reduction. This performance degradation is inherent in virtualization scenarios regardless of whether a single virtual machine or multiple virtual machines are running concurrently. The additional layer of abstraction introduced by the hypervisor and the sharing of physical resources among VMs inevitably result in decreased performance compared to running on dedicated physical hardware.

2. RESEARCH OBJECTIVES, MOTIVATION, AND GOALS

This paper focuses on comparing FS performance between hypervisor-based virtualization and native operating systems running on physical machines. In related studies, numerous papers have addressed similar issues, investigating the performance of different hypervisors, including VMware ESXi, KVM, Xen, Proxmox, and Hyper-V. Such studies rely on diverse hardware configurations and make use of established benchmarking tools like Filebench, Fio, Bonnie++, Postmark, HD TunePro, Iozone, and LMBench [2-8]. A majority of sources in the literature do not employ a mathematical model; instead, they conduct high-quality experiments on specific hardware setups. The findings from these experiments are highly applicable in practical scenarios.

Several papers have investigated comparisons between native operating systems and virtualized operating systems [7-14]. While a minority of studies suggest minimal performance degradation [7-8], the majority indicate significant drops in performance, aligning with our findings as presented in our paper. Our approach shares similarities with references [9-14] but distinguishes itself through an expanded mathematical model and extensive experiments. These experiments include additional workloads, a larger number of virtual machines (VMs), and unique results across diverse hardware configurations.

The primary goal of this paper is to assess the FS efficiency differences between a native operating system and a virtualized operating system utilizing the KVM

hypervisor (type-1). The study uses identical hardware configurations for both the native host operating system and the hypervisor with its configured virtual machines. Filebench is employed as the tool for performance measurement [15]. Central to this research is a foundational mathematical model, pivotal in interpreting the test results. This model is designed to be scalable, versatile, and applicable to similar case studies. The main focus lies in evaluating how virtualization impacts FS performance, with a specific emphasis on comparing the performance of the native operating system against that of the hypervisor running a single virtual machine. Additionally, the experiment is expanded by increasing the number of virtual machines from 1 to 4. The hypothesis that the native operating system outperforms KVM virtual machines was confirmed by the experiment's findings. This research contributes significantly by providing insights into the performance degradation associated with virtualization, supported by a robust mathematical framework.

3. KVM

KVM (Kernel-based Virtual Machine) technology holds a crucial role in Linux-based virtualization. Originally introduced with backing from Red Hat, KVM has been an integral part of Linux's core functionality starting from version 2.6.20. It operates as a kernel module, blurring the distinction between type-1 and type-2 hypervisors. KVM extends the capabilities of Linux's core architecture by enabling it to function as a native hypervisor. This setup allows Linux to host virtual machines directly, leveraging its own capabilities. Unlike standalone hypervisors, KVM doesn't require additional software like QEMU to manage VMs; instead, it utilizes existing Linux functionalities. As a hosted hypervisor, KVM leverages Linux as both the host operating system and the hypervisor, effectively embedding virtualization features within the Linux framework. In this hosted hypervisor setup, KVM utilizes QEMU [16] hardware virtualization without relying on its own.

4. MATHEMATICAL MODEL AND HYPOTHESES REGARDING EXPECTED BEHAVIOUR

In this paper, we look into the performance relation between native host operating systems and identical operating systems deployed through hypervisors. Our objective is to establish a model that evaluates FS performance across physical architectures and hypervisor



environments. In evaluating FS performance, benchmark tools are employed to generate specific workloads. These benchmarks typically encompass four main types of cycles: random reading, random writing, sequential reading, and sequential writing. Additionally, write operations can be categorized as synchronous or asynchronous, reflecting the significant influence of file system caching on write performance. Each benchmark workload simulates a diverse range of operations within the file system. These operations include managing directories, handling metadata, managing free lists, manipulating file blocks, and performing various housekeeping and journaling operations.

For modeling traditional architecture, we noticed there are the physical hardware and the host operating system (with the kernel and the file systems). The model involves three objects: a benchmark, a host kernel, and a host file system, so it means that the whole data path is simple. In the data-path, the benchmark produced the requests to the host kernel, and then the kernel forwarded these requests to the host file system. Workload processing time, TW_{hostOS} , depends on the benchmark, host kernel, and host file system, Equation 1:

$$TW_{hostOS} = f(Bch, hFS)$$

Equation 1. Host OS workload processing time

In Equation 1, Bch is the benchmark processing, including the file sets with their own file set operations. On the basis of the file set operations, the benchmark produces the requests to the host kernel. The kernel processes the requests, and then the kernel produces the requests to the host file system; this processing in the host file system is denoted as the hFS . The processing of the host file system is highly complex, involving the characteristics of three objects: the host file system, the file system cache, and the physical disk drivers.

For hypervisor virtualization, the workload processing time is much more complex, while the data path in virtualization depends on the large number of components. These are the features of the following: file system types on both the guest and host sides, the guest and host file system caching, the virtual machine image file, hypervisor processing, and the hypervisor parameters. Data-path includes seven objects: benchmark, guest kernel, guest file system, virtual hardware, hypervisor, host kernel, and host file system. The model includes the three kernels, and two file systems with their own file system caches in the form of the FS-pair, and so it is a solidly complex data path, Equation 2.

$$TW_{hyp} = f(B, gFS, VH_{proc}, Hyp_{proc}, hFS)$$

Equation 2. Hypervisor virtualization workload processing time

The first component, B denotes the benchmark requests for the guest OS file system. The second component, gFS , marks the processing of the guest OS file system, correlated with the kernel of the guest OS. This component is quite similar to the 5th component, hFS , and these components support different file system types. The third component, VH_{proc} , is the actual processing of the virtual disk drivers. The fourth component, Hyp_{proc} , refers to the hypervisor processing time, during which the hypervisor receives requests from the virtual disk drivers and forwards them to the virtual machine image file (VMI) in the host file system. The fifth component, hFS , refers to the host OS file system time processing, which is closely integrated with the kernel of the host operating system. This component operates with a large VMI file. The second and the fifth components of Equation 2 are very correlated and must be considered an FS-pair, involving the complex relation of two FS caches.

Observing the data path in Equation 1 (three input factors) and Equation 2 (five input factors), the data path is far more complex in virtualization. In summary, while native host operating systems generally exhibit superior FS performance due to direct hardware access, significant performance degradation is expected in virtualized environments. This degradation can vary based on workload characteristics and a multitude of factors related to physical hardware, operating systems, and virtualization technologies.

5. TEST SETUP AND BENCHMARKING APPLICATION

We focus attention on a fair comparison of file system performance achieved through identical hardware, virtual machines, measurement methodologies, operating systems, and a standardized benchmarking program. We used the KVM virtual platforms: QEMU emulator version 8.1.2 (pve-qemu-kvm_8.1.2-4) on a Linux host OS: Debian 12 Bookworm, kernel: 6.5.11-4 / ext4, while the experiment was conducted on an HP server running CentOS Stream 9 as the native guest operating system. The HP server features the following configuration:

- CPU: Intel® Xeon® Silver 4116 CPU @ 2.10GHz



- **RAM:** 32GB DDR4 2400 MHz
- **Hard disk:** 2x HPE 480GB RAID1, SATA 3, Sequential read up to 535 MB/s, Sequential write up to 495 MB/s
- **Host Operating Systems:** KVM: QEMU version 8.1.2 on Debian 12 Bookworm, kernel: 6.5.11-4, ext4

Each experimental test utilized Filebench 1.4.9.1-3 as the benchmarking tool. Filebench facilitates the simulation of diverse server environments by defining various workloads, offering detailed performance metrics such as file read/write throughputs [16]. The storage setup comprised two identical hard drives configured in RAID-1, housed within the HPE ProLiant BL460 Gen10 server. Both native and virtual environments were tested, with virtual machines stored on the same RAID-1 disks. Below are the parameters used for the virtual machines:

- **Number of virtual CPU assigned to each VM:** 4, Virtual memory assigned to each VM: 8GB
- **Virtual hard disk assigned to each VM:** 64GB (/dev/sda), 32GB root FS, 32GB testing FS (XFS)
- **Guest OS:** CentOS Stream 9

All performance tests were conducted using Filebench, a well-established file system and storage benchmarking tool. Filebench is renowned for its ability to simulate a wide range of workloads that closely resemble real-world server environments. These workloads can mimic services such as mail servers, web servers, database servers, file servers, and more. This capability allows for comprehensive testing and evaluation of file system and storage performance under conditions that mirror practical server usage scenarios.

Mail server results for the native host OS and for the virtual machines are shown in Figure 1.

Features of the mail server workload are the following: dominant random reading and random writing (synchronous), without sequential components. Other features are a moderate dataflow and a moderate number of input/output requests. As a consequence of predominant random reading and synchronous random writing, the caches of both file systems have a small performance influence. For the KVM mail server workload, watching the native host operating performance related to a single virtual machine (1VM), we detected remarkable performance degradation (about 2.73 times). As the number of virtual machines (1VM-4VM) increases, we determined the further performance drops (about 27-67%). For the KVM mail server workload, all speeds are solidly lower than the max disk speed (500MB/s). It means that FS cache effects are almost zero. For the KVM mail server, the differences between native and virtual operating systems are a consequence of Equation 1 vs. Equation 2. For a small dataflow and predominant synchronous random writing, the FS caches have no impact, so these differences are remarkable. For virtualization, in the view of Equation 2 and KVM mail server, we consider that two components have the dominant performance influence: the VH_{proc} component (Equation 2) above all and then the Hyp_{proc} (Equation 2), while some small influence is due to the FS components (gFS and hFS) (Equation 2), which operate as an FS-pair, exclusively. For the KVM mail server workload, the impact of two file system caches (as part of an FS-pair) is insignificant; the main reason is the dominant synchronous random writing/random reading.

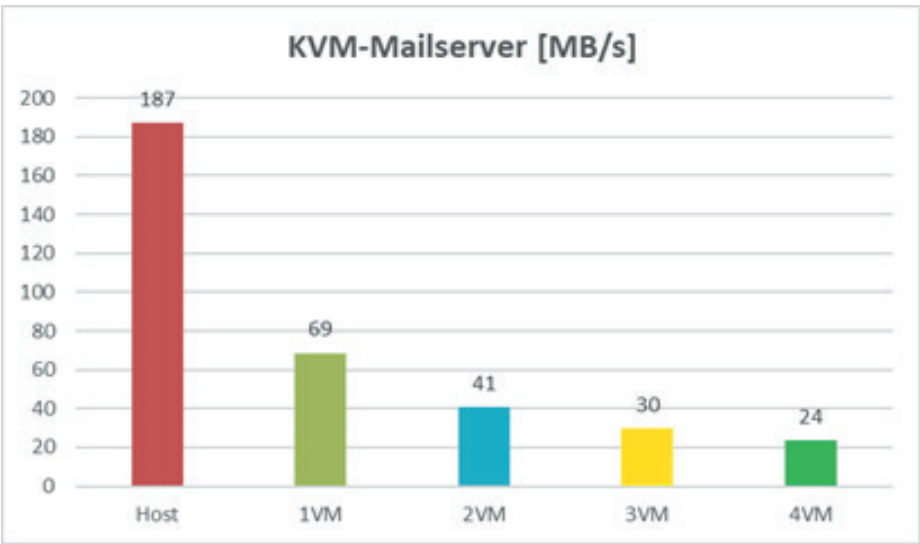


Figure 1. Mail server performance test results



Results for the web server for the native host OS and the virtual machines are shown in Figure 2.

Features of the web server workload are the following: a lot of random read components and small random write components (as the log-appending), whereas sequential transfers do not exist. Other features are the small dataflow and a moderate number of input/output operations. By the predominant random read components, the file system caches have inconsiderable performance influence unless the readings are with repetition. For the KVM web server workload, regarding the native host FS performance related to a single virtual machine, we noticed a relatively small performance drop (6%). With increasing the VM number (1VM-4VM), we detected the additional performance degradation (5%-26%). For the KVM web server workload, all speeds are lower than maximal disk speeds. Anyway, the speeds are high for random read workload, which means the repeated random reads exist, so the cache effect is noticed, but the random read data traffic for virtual/physical drivers is dominated, also. For the KVM web server, the speed differences between native and virtual operating systems are a consequence of Equation 1 vs. Equation 2. For a small dataflow with repeating random reading, the FS caches relieve these differences remarkably. Seen through the virtualization, in the view of Equation 2 and KVM web server, we consider that two components have the most performance impact. These are the Hy_{proc} (Equation 2) and the VH_{proc} (Equation 2), whereas the certain influence is by the FS components gFS , and hFS (Equation 2), in the form of the FS-pair. For the KVM web server workload, the influence of two file system caches (in an FS-pair) is solid; the reason is the dominant repeated random reading.

Fileserver results for the native host OS and the virtual machines are shown in Figure 3.

Features of the fileserver workload are the following: the dominant random/sequential reading and random and random/sequential write components; in other words, all kinds of transfers are present. Other features are the large dataflow and the large number of input/output requests. By the repeated reading and lots of writing, the file system caches have a solid performance influence. For fileserver workload, observing the native FS performance related to a single virtual machine, we noticed large performance drops (2.72 times). By increasing the number of VMs (1VM-4VM), we detected the additional performance degradation (11%-36%). For the KVM fileserver workload, the speeds are higher than the maximal disk speed (500MB/s) for all virtual machines. Anyway, the speeds are high for this kind of workload. It means that the FS cache impact is large for whole writing and for repeated reading, but the disk traffic for virtual/physical drivers is significant, also. For the KVM fileserver, all differences between native and virtual operating systems are a consequence of Equation 1 vs. Equation 2. Despite the large cache effects, the differences are still large. Seen through the virtualization, in the view of Equation 2 and the KVM fileserver, we identify two components as having the most significant impact on performance influence, the VH_{proc} and Hy_{proc} components. The big influence is due to the components gFS and hFS (Equation 2), which operate as an FS-pair also. For the KVM fileserver workload, the influence of two file system caches is very remarkable; the main reason is the multiple repeated readings and a solid amount of writing.



Figure 2. Web server performance test results

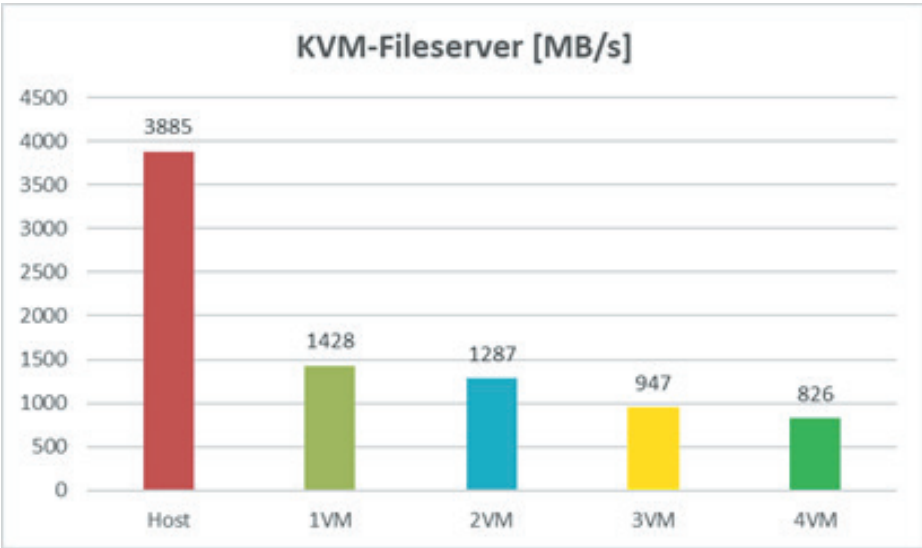


Figure 3. Fileserver performance test results



Figure 4. RFA performance test results

RFA results for the native host OS and the virtual machines are shown in Figure 4.

Features of the RFA workload are the following: the dominant random read components as well as asynchronous random write components; sequential transfers are very little present. Other features are the large number of input/output requests and the moderate dataflow. As a consequence of the dominant random asynchronous writes, the file system caches have a huge performance influence. For the KVM RFA workload, in the context of the native host performance versus a single virtual machine, we measured the small performance drop (20%). With increasing the number of virtual machines (1VM-4VM), we noticed the additional performance

drops (17-27%). For the KVM RFA workload, all RFA speeds are solidly higher than the max disk speeds; it means that the FS cache impact is extremely large. For the KVM RFA, the differences between native and virtual operating systems are due to Equation 1 vs. Equation 2. However, for a repeating random reading and asynchronous random writing, the FS caches reduced these differences. Seen through the virtualization, in the view of Equation 2 and KVM, we consider that two components have the most performance influence, the Hyp_{proc} and VH_{proc} components (Equation 2). For the KVM RFA workload, the influence of two file system caches is huge; the main reason is the prominent asynchronous random writing.

When we collect all results together, we can see the overall picture of this case study, as shown in Table 1.



Table 1. Overall results

%	Host/1VM	1VM/2VM	2VM/3VM	3VM/4VM
Mail server	2.73 times	67%	37%	27%
Web server	6%	5%	5%	26%
FS	2.72 times	11%	36%	15%
RFA	20%	20%	17%	27%

For this experiment, watching the native host FS performance related to a single virtual machine, we noticed solid drops in file system efficiency for all or most workloads. We detected the most pronounced drop (2.7 times) for the most complex workload, fileserver (with large dataflow), and 2.7 times for mail server (with smaller data flow). However, for less demanding workloads (web server and RFA) we have detected a smaller drop, 20% for RFA and 6% for web server. With increasing the number of virtual machines (1VM-4VM), solid drops are for fileserver 10-36% and for mail server 27%-67%. Then, we detected relatively smaller drops: for web server 5-26% and for RFA 17-27%. These relatively small drops with the increasing number of virtual machines are a consequence of FS caches for a strong physical server and a relatively large amount of RAM allocated to virtual machines.

6. CONCLUSION

In this paper, if we are looking at the native host FS performance related to a single virtual, as the main evaluation parameter, we detected significant drops in file system efficiency, with the most pronounced drop being about 2.7 times. This drop was detected for the most complex workload, which contains a dominant data flow, and for mail server workload. Drops for web server 6% and for RFA 20% represent the smaller drops. With further increasing the number of virtual machines (1VM-4VM for our case), the FS performance drops can be relatively strong, for fileserver 10-36% and mail server 27%-67% workloads. For web server the performance drops continue (5-26%) and for RFA 17-27%, and these can be seen as the relatively smaller FS drops. Anyway, we consider that KVM hypervisor-based virtualization exposes the solid drops in file system efficiency.

For a comprehensive exploration of the FS performance relationship between native host operating systems and hypervisor-based virtualization, our mathematical model is highly adaptable. However, generating numerous case studies is essential.

These studies will contribute to building a Knowledge Data Base (KDB) focused on understanding FS performance degradation caused by virtualization. The case studies will encompass various aspects: different hardware configurations, various hypervisors, diverse operating systems and their kernels, different file systems, various benchmarks, etc. This effort represents a significant component of our future research agenda.

7. ACKNOWLEDGMENT

The paper has been funded by the Ministry of Education, Science, and Technological Development of the Republic of Serbia.

REFERENCES

- [1] E. Correia, *Hypervisor based server virtualization*. Encyclopedia of Information Science and Technology, Third Edition, IGI Global, pp 1182-1187, doi:10.4018/978-1-4666-5888-2.ch112, 2015.
- [2] M. Polenov, V. Guzik, V. Lukyanov, *Hypervisors comparison and their performance*, CSOC2018: Software Engineering and Algorithms in Intelligent Systems, Springer, pp148-157. doi:10.1007/978-3-319-91186-1, 2018.
- [3] H. Kazan, L. Perneel, M. Timmermann, "Benchmarking the performance of Microsoft Hyper-V server, VMWare ESXi and Xen hypervisors", *J. of Emerging Trends in Computing and Information Sciences*, vol. 4, no. 12, pp. 922-933, 2013.
- [4] S. Pawar, S. Singh, "Performance comparison of VMWare and Xen hypervisor on guest OS", *IJICSE*, vol. 2, no. 3, pp. 56-60, 2015. <https://ijicse.in/index.php/ijicse/article/view/43/41>
- [5] S. A. Algarni, M. R. Ikbali, R. Alrooba, A. S. Ghiduk, F. Nadeem, "Performance evaluation of Xen, KVM, and Proxmox hypervisors", *Int. J. of Open Source Software and Processes*, vol. 9, no. 2, doi: 10.4018/IJOSSP.2018040103, 2018.



- [6] P. Kedia, R. Nagpal, "Performance evaluation of virtual environment with respect to physical environment", *Int. J. of Computer Applications*, 89(11), pp 17-22, doi: 10.5120/15676-4425, 2014.
- [7] V. K. Manik, D. Arora, "Performance comparison of commercial VMM: ESXi, XEN, HYPER-V & KVM", in *3rd Int. Conf. on Computing for Sustainable Global Development, New Delhi*, <https://ieeexplore.ieee.org/document/7724572>, 2016.
- [8] A. Bhatia, G. Bhattal, "A comparative study of various hypervisors performance", *Int. J. of Sci, Eng. and Tech. Research*, vol. 7, no. 12, pp. 65-71, <https://www.ijser.org/researchpaper/A-comparative-study-of-Various-Hypervisors-Performance.pdf>, 2016.
- [9] B. Đorđević, V. Timčenko, E. Nikolić, N. Davidović, "Comparing Performances of Native Host and Virtualization on ESXi hypervisor", *IEEE in 20th INFOTEH-JAHORINA*, 21(1), pp 1-4, doi: 10.1109/INFOTEH51037.2021.9400648, 2021.
- [10] B. Đorđević, S. Milenković, N. Davidović, V. Timčenko, "Performance comparison of native host vs. ESXi hypervisor-based virtualization", *8th IcETRAN*, https://www.etrn.rs/2021/zbornik/Papers/101_RTI_2.2.pdf, 2021.
- [11] B. DJORDJEVIC, V. TIMCENKO, N. KRALJEVIC, N. MACEK, *File System Performance Comparison in Full Hardware Virtualization with ESXi, KVM, Hyper-V and Xen Hypervisors*, *Advances in Electrical and Computer Engineering (AECE)*, 21(1), pp 11-20, doi: 10.4316/AECE.2021.01002, 2021.
- [12] B. Đorđević, V. Timčenko, D. Sakić, N. Davidović, "File system performance for type-1 hypervisors on the Xen and VMware ESXi", *IEEE in 21st INFOTEH-JAHORINA*, doi: 10.1109/INFOTEH53737.2022.9751288, 2022.
- [13] B. Đorđević, M. Marjanović, V. Timčenko, "Performance comparison of native host and hyper-based KVM virtualization", in *28th TELFOR*, doi: 10.1109/TELFOR51502.2020.9306550, 2020.
- [14] B. Đorđević, M. Piljić, N. Kraljević, V. Timčenko, "Comparison of file system performance in full virtualization with MS Hyper-V and KVM hypervisors", in *9th IcETRAN*, https://www.etrn.rs/2022/zbornik/ICETRAN-22_radovi/067-RTI2.5.pdf, 2022.
- [15] Filebench - <https://github.com/filebench/filebench>. [Accessed 2023].
- [16] H. D. Chirammal, A. Mukhedkar, A. Vettathu, *Mastering KVM Virtualization*, Packt Publishing Ltd, ISBN 9781784399054, 2016.



FILE SYSTEM PERFORMANCE COMPARISON WITH KVM AND XEN AS TYPE-1 LINUX-BASED HYPERVISORS

Borislav Đorđević*,
[0000-0002-6145-4490]

Kristina Janjić,
[0009-0005-0750-6105]

Nenad Kraljević
[0009-0008-7684-5444]

School of Electrical and Computer
Engineering,
Belgrade, Serbia

Abstract:

This study compares FS (file system performance between bare-metal hypervisors built on the Linux architecture, focusing on KVM and Xen. While both divide Linux-based and QEMU-based architectures, detailed analysis reveals important differences. Filebench was chosen as the benchmarking tool for its adaptability in emulating real-world applications within authentic server environments. CentOS 9, a representative Linux distribution, served as the guest OS. Performance was assessed while running one, two, and three VM (virtual machine) concurrently, highlighting scalability under varying loads. The study introduces a mathematical model of the bare-metal virtualized environment to establish a theoretical framework for performance analysis. Empirical experiments complement the model, serving as specific case studies. By combining theoretical modeling with practical experimentation, the research provides a deeper understanding of the factors influencing FS performance in virtualized environments.

Keywords:

Hypervisor, KVM, Xen, CentOS 9, Virtual Machine - VM, Filebench.

INTRODUCTION

Virtualization is a transformative technology in contemporary IT, fundamentally altering the way resources are handled, stored, and utilized. By enabling several OS (operating system) to coexist on the same HW (hardware) platform, it enhances system reliability, availability, and resource efficiency. Virtualization enables the creation of VM, servers, along with other HW resources, leading to cost reductions, simplified system management, and greater expandability. Through precise allocation of HW resources like processing power (CPU), RAM, and disk space to VM, it ensures optimal HW utilization [1]. Although virtualization provides many advantages, it also introduces challenges, such as complex management, security vulnerabilities, potential software licensing expenses, and the vulnerability of having a single point of failure in case of hypervisor or physical server issues. However, the benefits of virtualization greatly outweigh its drawbacks, establishing it as a cornerstone of contemporary IT.

Correspondence:

Borislav Đorđević

e-mail:

borislav.djordjevic@viser.edu.rs





Virtualization includes several forms, including HW, SW (software), data, and desktop, as well as the virtualization of storage, memory, and network HW configurations. This study concentrates on HW virtualization, which allows VM to run as independent computers while sharing the same physical HW. This capability is enabled by the hypervisor, which creates and manages VM. Hypervisors abstract the HW from the OS, allowing numerous OS to operate in parallel on a single HW platform. Hypervisors are typically categorized into two types: type-1 hypervisors, or bare-metal hypervisors, running directly on the HW, and type-2 hypervisors, also known as hosted hypervisors, functioning within an existing OS environment.

2. MOTIVATION, RESEARCH OBJECTIVES, AND PURPOSE

A variety of scholarly articles within the virtualization field employ different techniques to evaluate the efficiency of different virtual environments. One widely adopted approach is to conduct comparative analyses that examine the performance of different hypervisors [2-10]. These analyses usually consist of meticulously planned experiments and rely on widely recognized benchmarks. However, it is noteworthy that a large number of these studies omit the use of mathematical modeling when evaluating virtual environments.

References [11-15] employ analogous mathematical models in the present paper, yet they concentrate on distinct hypervisors and operate on diverse HW specifications.

The primary purpose of this study is to develop a thorough mathematical model aimed at evaluating FS performance in virtual environments using type-1 hypervisors built on the Linux architecture. This model incorporates a broad set of input arguments and is made to be expandable for future improvements. What differentiates our approach from others is the methodology we apply: we start by developing a mathematical model, followed by an exploratory setup that acts as a particular practical analysis. This math model plays a pivotal role in analyzing the outcomes of our experiment, providing a distinct viewpoint on the assessment of the performance of virtual environments.

In this study, we focused on comparing KVM and Xen as type-1 hypervisors, based on the Linux framework hypervisors, utilizing QEMU-based full HW virtualization technology. The same HW conditions were applied for testing both hypervisors to ensure

an impartial evaluation. The guest OS was CentOS 9, configured with the XFS FS. For the experiments, we employed the Filebench performance testing program, which includes different distinct workloads: fileserver, mail server, web server, and RFA (random-file-access). These workloads were selected to represent different types of typical virtualized environment tasks. Following the experiments, we validated our findings through the implementation of a mathematical model to analyze and clarify the results. This approach allowed us to derive meaningful insights into the performance characteristics of KVM and Xen across various workloads within the virtualized environment.

3. XEN AND KVM

The Xen platform consists of the Xen hypervisor, which runs directly on the physical HW, and multiple domains that function as VM operating atop the hypervisor. The main elements working together to provide efficient and flexible virtualization include:

- **Xen Hypervisor:** This is the core software layer that interacts directly with the HW. It is responsible for directing resources such as processing power, RAM, and input/output operations for multiple guest OS running simultaneously. The hypervisor ensures the parallel execution of these guest systems and supports various architectures, including x86, x86-64, Itanium, PowerPC, and ARM.
- **Domain 0 (Dom0):** Dom0 is a modified Linux kernel that is tightly integrated with the hypervisor. It holds exclusive rights for managing physical input/output resources and overseeing other VM (DomU guests). At startup, Dom0 is initiated and is tasked with both controlling the Xen hypervisor and running the device drivers for physical HW.
- **Domain U (DomU):** DomU refers to unprivileged guest systems that are not able to directly access physical HW. Instead, physical HW is managed by Dom0. These guests can be initiated through either a modified OS with paravirtualization or an OS that remains unchanged utilizing HW-assisted virtualization.

This hypervisor regulates access to the HW of the physical machine for guest domains. Instead of allowing guest domains to directly utilize privileged instructions, hypercalls are utilized to notify the Xen hypervisor of the need to execute privileged instructions, after which the hypervisor handles the request. The functionality of



hypercalls is analogous to system calls in an OS. It serves as a software catch between the VM and Xen, much like how a system call acts as a software interrupt between an application and the kernel [15].

KVM (Kernel-based Virtual Machine) is a virtualization solution that is open-source and enables HW-level virtualization directly integrated into Linux and functions as part of its kernel (Figure 3). KVM was originally developed as a Red Hat-sponsored initiative and has seamlessly integrated itself into the Linux kernel starting from the version 2.6.20, functioning as a crucial module of kernel.

KVM utilizes parts of QEMU for emulating real devices. One of KVM's advantages is its ability to support a diverse range of OS for VM, including both Linux and Windows. The KVM hypervisor enables full virtualization, utilizing HW virtualization on supported processors to allow VM to operate without modifications to the guest OS. It offers every VM the full range of services typically found in a physical system, including a virtual-BIOS and HW such as a processor, memory, storage, network cards, etc. As a result, each VM fully simulates a physical computer [16].

4. MATHEMATICAL MODEL AND ASSUMPTIONS ON EXPECTED BEHAVIOR

The time required to process workloads in a hypervisor-based virtual environment, Tw , is determined by a minimum of five key elements that exert significant influence, Equation 1:

$$Tw = f(Bn, gFS, VH-proc, Hyp-proc, hFS)$$

Equation 1. Wokload time in virtual environment

The initial element, Bn , depicts the time taken by the benchmark to process data. The second element, gFS , denotes the processing time of the FS within the guest OS. The guest FS is closely linked with several components, including the guest OS kernel and its file system cache mechanism. Both Bn , gFS expose comparable characteristics across all tested hypervisors. This similarity stems from the consistent use of the same benchmark features, parameters, VM configurations, and the guest FS (XFS).

$VH-proc$, the third element, depicts the processing time for virtual HW, primarily the virtual disk drivers. During our experiment, both KVM and Xen hypervisors show considerable differences in the $VH-proc$ component. KVM exclusively employs full HW virtualization, while

Xen supports both full HW virtualization and paravirtualization, although paravirtualization is not part of the scope of our experiment. It is important to note that both hypervisors rely on QEMU-based open-source technology for full HW virtualization. A key distinction, however, lies in the sets of virtual drivers provided by QEMU, which have evolved over a long period (from the release of qemu-0.10.0 on March 4, 2009, to the release of qemu-8.2.1 on January 29, 2024). Xen and KVM each incorporate different sets of QEMU virtual drivers, which significantly impact the $VH-proc$ component. As a result, the performance characteristics of each hypervisor are expected to differ substantially. $VH-proc$ is intricately linked with FS caching on both the guest/host OS sides. The differences in $VH-proc$ between Xen and KVM can therefore impact how effectively each hypervisor manages FS caching, further influencing overall system performance in our experimental setup.

$Hyp-proc$, the fourth element, depicts the time allocated for hypervisor processing. This includes the time the hypervisor spends handling requests from virtual drivers and forwarding them to the host OS. In practical terms, FS requests from the guest FS are passed on to the host FS within the VMI (VM image file). In our experiment, we evaluated two distinct hypervisors: XenServer with the original Xen hypervisor, which follows the traditional Xen architecture, and KVM hypervisors, which use real Linux kernels with KVM kernel modules. These hypervisors are designed with a micro-kernel architecture, which is both lightweight and modern. Despite both hypervisors employing a micro-kernel design, they are expected to exhibit different hypervisor processing times due to their unique design philosophies and implementation details. These differences in hypervisor processing times ($Hyp-proc$) are crucial as they directly impact the overall performance and responsiveness of virtualized environments. Understanding these nuances is essential for accurately interpreting and comparing the performance results obtained from experiments involving Xen and KVM hypervisors.

hFS , the fifth element, depicts the processing time of the host FS, and we expect notable variations among hypervisors. In our testing environment, both hypervisors, being Linux-based, use two widely used FS options: ext4 and XFS, which may or may not involve the use of the LVM (Logical Volume Manager). For our setup, we selected ext4 for both hypervisors, though it's important to mention that the versions of ext4 used were not identical, which leads to inherent differences in performance.



When analyzing the host OS as a key element of the virtual environment, several important differences can be observed between the two Linux-based hypervisors, Xen and KVM. Both hypervisors operate with a distinct Linux distribution: XenServer utilizes customized Linux version tailored for Xen, while KVM can work with any Linux distribution. As a result, the host OS differs in several aspects, such as versions of physical drivers, host kernels, versions of host FS, and OS-system/graphical environments. While both hypervisors are Linux-based, they exhibit considerable variations in their host OS configurations and FS.

In summary, the features regarding the similarities and differences between the two hypervisors are as follows:

On the guest side, all features of the VM and benchmark remain the same.

On the virtual environment side, both hypervisors utilize full HW virtualization and are based on QEMU solutions for virtual drivers. However, significant differences arise due to the third (*VH-proc*) and fourth (*Hyp-proc*) components in Equation 1.

On the host OS side, despite both hypervisors being Linux-based, there are substantial differences in kernel versions, Linux distributions, host FS (both hypervisors use ext4, but with different versions), physical disk drivers, and FS cache mechanisms. In a virtual environment, an FS pair always exists (guest FS on host FS). Our experiment includes only one FS pair: XFS on ext4.

Using the mathematical model, along with the recognized similarities and distinctions between the hypervisors tested, we analyze and interpret the performance outcomes from the experiment.

5. TEST SETUP AND BENCHMARKING PROCESS

We emphasize fair and accurate performance evaluation by employing identical HW, VM, measurement techniques, OS, and a unified benchmarking tool. Fairness was maintained throughout the experiment by configuring consistent HW configurations, selecting a uniform OS for both the guest and host environments, and using a single benchmarking program across all stages of testing. The virtual environments employed were VMware ESXi version 8.0 and Xen Citrix Hypervisor version 8.2.1. The tests were conducted on an HP server running CentOS Stream 9 as the guest OS. The server's specifications are as follows: Intel® Xeon® Silver 4116 CPU @ 2.10GHz, 32GB DDR4 2400 MHz RAM, 2x

HPE 480GB SATA 6G RI SSF SSDs in RAID1 configuration with sequential read speeds up to 535 MB/s and sequential write speeds up to 495 MB/s, running Xen Citrix Hypervisor 8.2.1 and KVM with QEMU emulator version 8.1.2 (pve-qemu-kvm_8.1.2-4) on a Linux host OS (Debian 12 Bookworm, kernel 6.5.11-4, ext4).

All tests in the experiment were carried out using the Filebench 1.4.9.1-3 benchmark tool, which facilitates the reproduction of different real server scenarios by creating diverse workloads. It provides comprehensive performance data, including file read/write throughput for various types of workloads [17]. For storage, a pair of identical hard drives was set up in RAID-1 and mounted on the server (HPE ProLiant BL460 Gen10). Each virtual environment was tested with VM hosted on the same RAID-1 disks. The VM specifications are as follows: 4 virtual CPUs, 8GB of virtual memory, a 64GB virtual hard disk with 32GB allocated to /dev/sda1 (root FS) and 32GB to /dev/sda2 (testing FS with XFS), and the guest operating system is CentOS Stream 9.

6. EVALUATION AND FINDINGS

We single out that the primary goal of this study is to assess the FS performance of two dissimilar type-1 hypervisors using a range of workloads, including mail server, web server, fileserver, and random-file-access. Originally, performance was assessed with a single VM, followed by constant assessments with two, three, and four VM functioning concurrently.

The explanation of performance heavily relies on the features discussed in Chapter 4, including *VH-proc*, *Hyp-proc*, guest FS, FS-pair, FS-cache-pair, physical and virtual disk drivers, and various elements of the host OS, such as the kernel, host FS, and OS/graphical environments. It's crucial to emphasize that many of these features differ significantly between Xen and KVM.

Figure 1 illustrates the outcomes of the fileserver workload test.

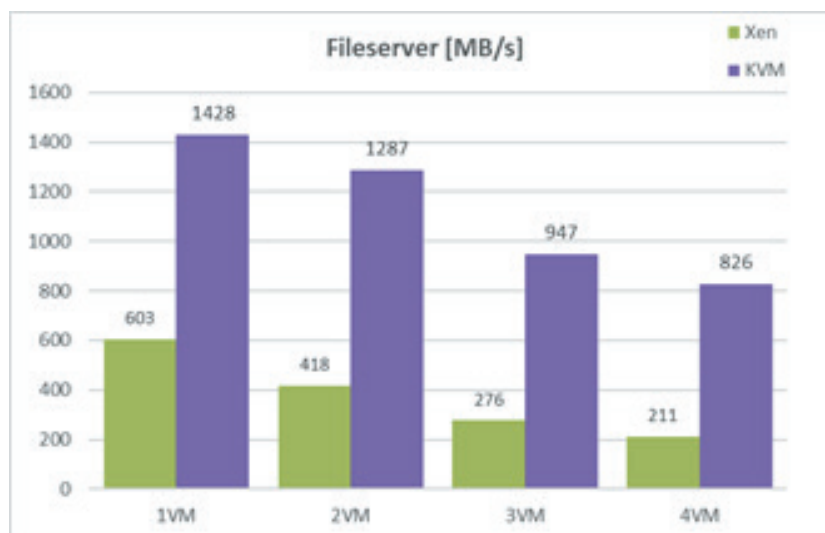


Figure 1. Results from the fileserver tests

The fileserver workload is defined by a rich mix of random and sequential read/write operations. This type of workload involves numerous input/output requests and a significant data throughput. FS caches play a crucial role, especially for frequent read operations and asynchronous writes. When examining the fileserver workload, KVM consistently outperforms Xen, with KVM being 2.37 to 3.92 times faster. The differences in FS performance are substantial. When observing the achieved throughput and the maximum disk speeds (around 500MB/s), both hypervisors surpass the maximum disk speeds with one VM. However, when using two VM, Xen's throughput drops below the maximum disk speeds, while KVM maintains higher throughput across all VM. These high throughputs indicate that FS cache pairs (guest/host caches) have a significant impact, while much of the I/O traffic also interacts with virtual-physical drivers due to the random and sequential nature of the workload. In the case of the fileserver workload, each element from Chapter 4 plays a role, but the most critical elements are *VH-proc* (including FS-pairs and drivers), disk drivers, and the cache effects of the FS-cache pair. Cache misses in both the guest and host caches make the virtual disk drivers of the guest OS and the physical disk drivers of the host OS very important factors. Given the substantial data throughput of the fileserver workload, it can be concluded that KVM benefits from a more efficient combination of components: *VH-proc* and the FS cache effects from the FS-cache pair, along with superior virtual/physical disk driver performance compared to Xen for random and sequential I/O requests.

Figure 2 illustrates the outcomes of the mail server workload test.

The mail server workload is mainly defined by a higher frequency of random reads and synchronous random writes, accompanied by a moderate volume of input/output operations and data throughput. Due to the prevalence of random reads and synchronous writes, the impact of FS caches on both the guest and host OS is anticipated to be minimal. For the mail server workload, Xen outperforms KVM, showing a performance improvement of 16-30-62%. When observing the achieved speeds and the maximum disk speeds, both Xen and KVM exhibit notably lower throughputs across all VM. The low mail server speeds suggest that the influence of guest/host FS caches is minimal, meaning most I/O traffic is directed to virtual/physical drivers. For the mail server workload, several components from Chapter 4 play a significant role. However, the primary components are *Hyp-proc* and *VH-proc* with FS-pair, although the FS cache effects are minimal in this case. Due to the limited influence of both FS caches, the virtual drivers of the guest OS and the physical disk drivers of the host OS become the most critical factors, particularly for random read/random write traffic. Given this, we conclude that Xen provides a more optimal combination of *Hyp-proc*, *VH-proc*, disk drivers, and minimal cache effects for the mail server workload.

Figure 3 illustrates the outcomes of the web server workload test.

The web server workload is defined by a prevalence of random reads and small random writes, accompanied by a fair volume of input/output requests and data throughput.

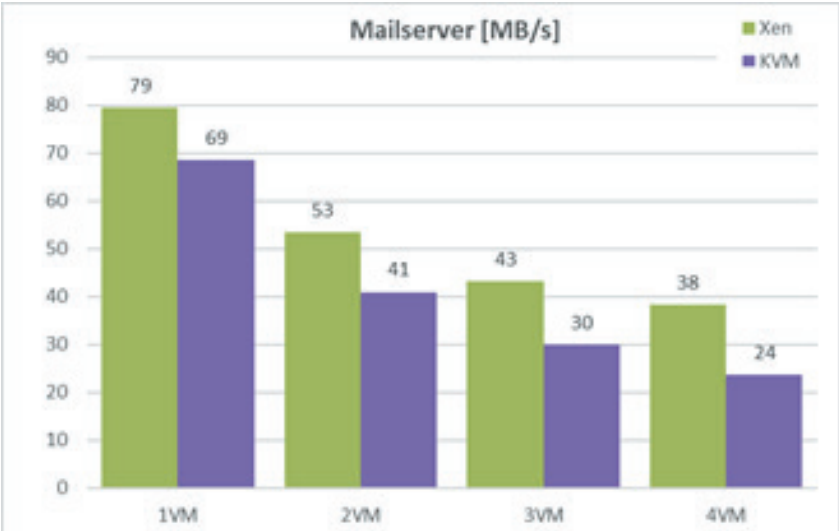


Figure 2. Results from the mail server tests

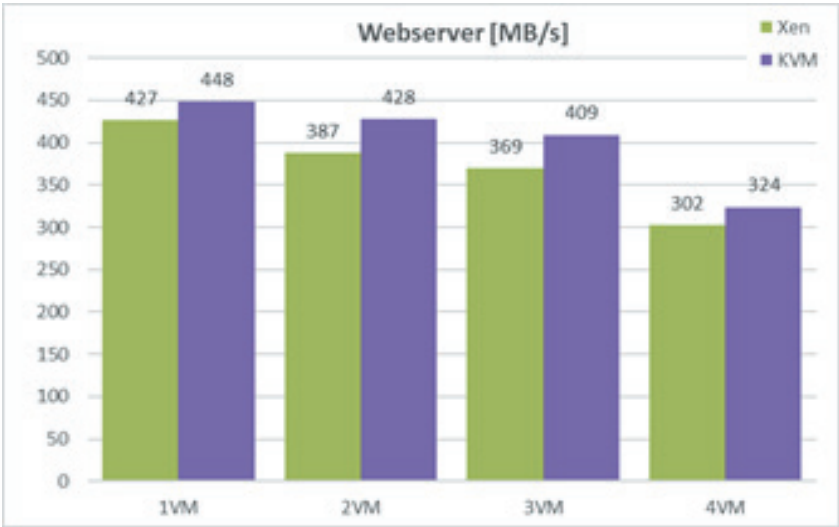


Figure 3. Results from the web server tests

The impact of FS caches on both the guest and host OS for random reads may be minimal, unless the reads are repetitive. For the web server workload, KVM slightly outperforms Xen, with a performance improvement of 5-10%. Observing the achieved web server throughputs and the peak disk speeds reached, both KVM and Xen exhibit relatively high random-read speeds, indicating a solid impact of FS caches on random reads. However, much of the random I/O traffic is still directed to virtual/physical drivers. Towards this type of random read workload, all elements discussed in Chapter 4 are significant. However, the primary components are *Hyp-proc*, *VH-proc* with FS-pair, and the significant cache effects from the FS-cache-pair. Given the numerous cache misses in the two FS caches, the physical disk

drivers (host OS) and virtual drivers (guest OS) play a critical role. In the context of the web server workload, we presume that KVM and Xen exhibit similar combinations involving *VH-proc* with FS cache effects, *Hyp-proc*, and physical/virtual disk drivers, though KVM shows a slight advantage.

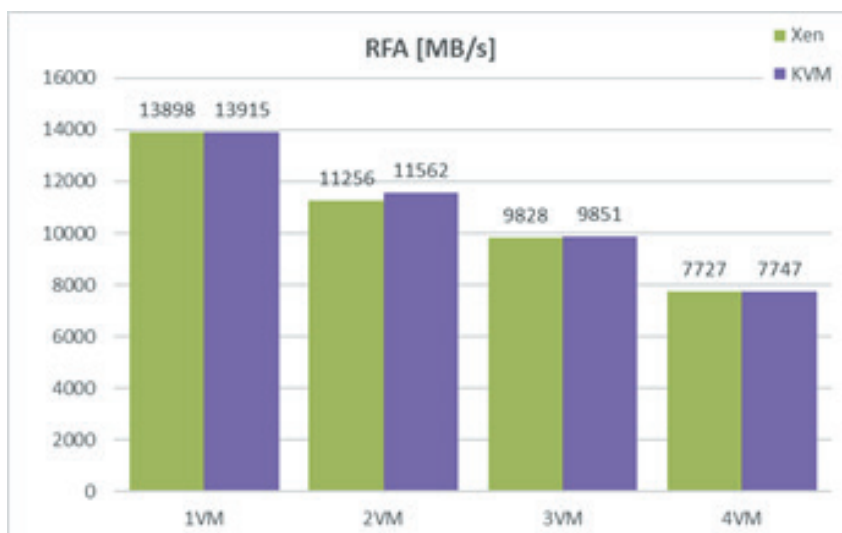


Figure 4. Results from the fileserver tests

Figure 4 illustrates the outcomes of the RFA workload test.

Primarily, the RFA workload is defined by random reads and asynchronous random writes, with a fair volume of input/output requests and a moderate data throughput. The presence of asynchronous writes enhances the importance of FS caches in this workload. In the case of the RFA workload, KVM slightly outperforms Xen, with a performance advantage of 0.1-3%. Observing the achieved RFA throughputs and the peak disk speeds, both of the hypervisors outperform the disk speeds notably. The high RFA throughputs suggest that the guest/host FS caches are the primary influencing factor. In the case of this random read/write workload, all the components discussed in Chapter 4 play an important role. However, we consider that the *VH-proc* component (with FS-pair) plays a dominant role, with exceptionally strong cache effects. In the RFA environment, both KVM and Xen demonstrate an excellent combination of *VH-proc* with strong FS cache effects, although Xen shows a slight edge.

CONCLUSION

We have evaluated the similarities and differences between the two Linux-based hypervisors, Xen and KVM. According to our mathematical model, the differences between these hypervisors stem from several key factors. While they may initially seem quite similar, a closer analysis reveals substantial differences in hypervisors like Xen and KVM. Despite sharing fundamental components such as *VH-proc*, *Hyp-proc*, and

hFS, their performance can differ significantly. In our study, the KVM hypervisor excels in fileserver, RFA, and web server workloads, while Xen performs better in mail server workload. These differences are substantial for fileserver workload (2.37-3.92 times), significant for mail server workload (16-62%), relatively minor for web server workload (5-10%), and slight for the RFA workload (0.1-3%).

To reach strong conclusions, it is crucial to perform multiple experiments across different case studies. We propose several potential avenues for future research, including the FS comparison of different type-1 Linux-based hypervisors under various HW configurations and workloads. This includes the analysis of upcoming hypervisor releases to understand any improvements or changes in their capabilities. Additionally, future research could explore various guest OS (including different versions of both Linux and Windows), investigate various FS such as ext4, XFS, and Btrfs, compare alternative benchmarking programs (HD Tune Pro, AS SSD, Fio), and conduct experiments to evaluate the influence of factors like RAM size and the number of CPU cores on performance.

ACKNOWLEDGEMENTS

The Ministry of Education, Science, and Technological Development of the Republic of Serbia has partially funded the work presented in this paper.



REFERENCES

- [1] E. Correia, *Hypervisor based server virtualization*, Encyclopedia of Information Science and Technology, Third Edition, IGI Global, 2015.
- [2] A. Kumar, S. Shiwani, "Guest operating system-based performance comparison of VMware & Xen hypervisor," *Int. J. of Sci, Eng. and Tech*, vol. 2, no. 5, pp. 286-297, 2014.
- [3] H. Kazan, L. Perneel, M. Timmermann, "Benchmarking the performance of Microsoft Hyper-V server, VMware ESXi and Xen hypervisors," *J. of Emerging Trends in Computing and Information Sciences*, vol. 4, no. 12, pp. 922-933, 2013.
- [4] A. Bhatia, G. Bhattal, "A comparative study of various hypervisors performance," *Int. J. of Sci, Eng. and Tech. Research*, vol. 7, no. 12, pp. 65-71, 2016.
- [5] S. Pawar, S. Singh, "Performance comparison of VMware and Xen hypervisor on guest OS," *IJICSE*, vol. 2, no. 3, pp. 56-60, 2015.
- [6] C. D. Graziano, *A performance analysis of Xen and KVM hypervisors for hosting the Xen worlds project*, Iowa: Iowa State University, 2011.
- [7] V. P. Singh, *Analysis of system performance using VMware ESXi server virtual machines*, PhD Thesis, 2012.
- [8] S. Ally, "Comparative analysis of Proxmox VE and XenServer as type 1 open source based hypervisors," *Int. J. of Sci. & Tech. Research*, vol. 7, pp. 72-77, 2018.
- [9] S. A. Algarni, M. R. Ikbāl, R. Alroobaea, A. S. Ghiduk, F. Nadeem, "Performance evaluation of Xen, KVM, and Proxmox hypervisors," *Int. J. of Open Source Software and Processes*, vol. 9, no. 2, 2018.
- [10] V. K. Manik, D. Arora, "Performance comparison of commercial VMM: ESXi, XEN, HYPER-V & KVM," in 3rd Int. Conf. on Computing for Sustainable Global Development, New Delhi, 2016.
- [11] B. Đorđević, M. Piljić, N. Kraljević, V. Timčenko, "Comparison of file system performance in full virtualization with MS Hyper-V and KVM hypervisors," in *Proceedings X ICETran*, Novi Pazar, 2022.
- [12] B. Đorđević, V. Timčenko, N. Kraljević, N. Maček: *File System Performance Comparison in Full Hardware Virtualization with ESXi, KVM, Hyper-V and Xen Hypervisors*, Advances in Electrical and Computer Engineering, 2021.
- [13] B. Djordjevic, V. Timčenko, N. Kraljevic, N. Jovicic, "Performance comparison of KVM and Proxmox Type-1 Hypervisors," in 30th TELFOR, Belgrade, 2022.
- [14] B. Đorđević, D. Sakić, N. Kraljević, V. Timčenko, N. Davidović, "File system performance for type-1 hypervisors on the Xen and VMware ESXi," in 21st INFOTEH-JAHORINA, East Sarajevo, 2022.
- [15] XenServer - <https://docs.xenserver.com/en-us/citrix-hypervisor>. [Accessed 2023].
- [16] H.D. Chirammal, P. Mukhedkar, A. Vettathu, *Mastering KVM Virtualization*, Packt Publishing Ltd, ISBN 9781784399054, 2016.
- [17] Filebench - <https://github.com/filebench/filebench>. [Accessed 2023].



THE PROBLEMS AND SOLUTIONS DURING AGILE TEAMS MEETINGS IN IT COMPANIES – A LITERATURE REVIEW

Maja Gaborov^{1*},
[0000-0002-3810-6156]

Nada Ratković Kovačević²,
[0000-0001-6398-4391]

Dragan Kreculj²,
[0000-0003-3268-4024]

Mihajlo Mulić³
[0009-0004-7448-0286]

¹University of Novi Sad,
Technical Faculty "Mihajlo Pupin",
Novi Sad, Serbia

²The Academy of Applied Studies
Polytechnic,
Belgrade, Serbia

³Alianza, inc,
Novi Sad, Serbia

Abstract:

The paper gives a literature review regarding the problems faced by IT (Information Technologies) companies during meetings of agile teams. The goal of the paper is to identify existing issues and propose potential solutions. Existing problems identified through the literature review were: lack of focus, engagement and transparency; technological difficulties, and conflicts in the team. Lack of focus is often the result of poorly structured meetings without defined goals, and lack of engagement leads to poor communication and unproductive discussions. Technological issues can disrupt the flow of meetings, and a lack of transparency can lead to team confusion. Conflicts in the team negatively affect the atmosphere and productivity.

A literature search was performed using selected academic databases. Abstracts of identified references were reviewed to assess their relevance. Selected studies were analysed to extract data on agile meeting challenges and solutions. Findings from different sources were synthesized to develop a comprehensive overview of issues and corresponding solutions. Suggested solutions include: conducting short, well-structured meetings; encouraging active participation of all team members; using stable technological solutions; regularly establishing open communication channels for conflict resolution. The results obtained indicate the need to improve agile meetings and consequently to enhance the performance of agile teams.

Keywords:

Agile Meetings, Agile Teams, IT (Information Technologies) Companies, Problems, Solutions.

INTRODUCTION

Agile methods have become crucial in software development due to their flexibility and focus on rapid adaptation to changes [1]. Agile teams are self-organized, cross-functional teams that work together to deliver value through iterative development [2]. Agile meetings are regular meetings that allow teams to discuss progress, challenges, and plans [3]. They promote communication and continuous improvement. Agile meetings are important for maintaining team cohesion and productivity.

The goal of this paper is to present the overview of the team problems in agile meetings and to suggest potential solutions relevant for IT companies which were found through a literature review.

Correspondence:

Maja Gaborov

e-mail:

maja.gaborov@tfzr.rs





Agile meetings often run into problems that can compromise team effectiveness [4]. Agile team meeting issues are extremely important for IT (Information Technologies) companies. Meetings must be effective because they are the basis for coordinating activities in agile teams. Due to the ineffectiveness of the meetings, there may be a loss of time and resources, which can then affect the development of the project [4]. Effective communication is key for teams. Difficult communication and communication problems can lead to misunderstandings, errors in work, and ultimately to a drop in the company's productivity [5]. It is desirable that the team members as well as teams cooperate with each other and support each other. If there are problems in the meetings, it can affect the team dynamics making it bad or disrupting it which can ultimately diminish team motivation [6].

It is necessary for the meetings to be productive, to identify problems and to respond to those problems in a timely manner in order to prevent or alleviate delays in the realization and delivery of the project [7]. Lack of clear guidelines and inconsistency among team members can lead to mistakes that could have been avoided with more efficient planning of meetings [8]. Poor communication in meetings can lead to conflict among team members, which can create a bad atmosphere in the working environment [9].

Effective meetings also assume that clients are timely informed about the progress of the project and that their feedback is taken into account, which would increase their satisfaction. Addressing these issues can improve the performance of agile teams, and thus the overall success of an IT company [10].

2. METHODOLOGY

The aim of this paper is to present team problems, during agile meetings in IT companies. After that, solutions to the problems are also explored and suggested. In order to find problems, and potential adequate solutions, literature search methodology is applied [11].

Based on the research objective, the research questions we defined are:

- RQ1: What are the common problems agile teams face during agile meetings?
- RQ2: What potential solutions can be proposed to solve these identified problems in agile meetings?

A literature search was performed using the following academic databases: Google Scholar, IEEE Xplore, ACM Digital Library, Springer, Wiley Online Libraries. A combination of keywords such as "agile meetings", "agile team questions", "focus on agile meetings", "engagement in agile teams", "technology issues in meetings", "transparency in agile teams" and "team conflicts in agile development" was used. The books, studies and articles published over the last 20 years were included to be investigated further, with an emphasis on those not later than the last 10 years, which deal with problems and solutions in agile teams and meetings. Papers that did not directly engage in the exploration of agile methodologies, or did not provide empirical data or articles that were not peer-reviewed were excluded. The abstracts of identified papers and other sources were reviewed to evaluate relevance. Relevant papers were selected and detailed information on the problems and proposed solutions gathered. Findings from various sources were organized and synthesized in order to include all problems and corresponding solutions in the review paper.

3. IDENTIFICATION AND OVERVIEW OF CHALLENGES IN AGILE TEAMS

Agile meetings are designed as efficient, focused and collaborative events. They are specified by emphasizing dynamic communication and adaptability. In agile teams, it's easy for participants to drift off topic, which can waste valuable time and reduce the overall effectiveness of the meeting. Lack of focus can be result of poorly structured meetings without clear objectives. In order to solve this issue, it is recommended to conduct well-organized meetings with defined goals and time limits. The approach like this can help ensure that the team remains focused on the task and maximizes the use of meeting time [12], [4].

Lack of team member engagement can lead to ineffective communication and unproductive conversations. When team members are not actively participating, the meeting may not achieve the desired results. It is necessary to foster a way of conduct and an environment in which each team member is encouraged to actively participate. This can be achieved by involving team members more e.g. through questions and by encouraging them to contribute to discussions. In this manner improving overall engagement and productivity would be achieved [13], [14].



Problems such as a poor Internet connection or inadequate video conferencing tools can disrupt the flow of a meeting, in case the hybrid form is chosen, making it difficult for participants to communicate effectively. These disruptions can interfere with a team's ability to collaborate and make decisions. It is important to use reliable video conferencing tools and have options to resolve any technical difficulties that may arise during the meeting [15], [16].

A lack of transparency can lead to confusion and reduced team productivity, as it may not be clear to team members what is expected of them. Sharing information about progress and task priorities can help ensure that all team members are on the same page and understand the current state of the project being developed [17], [18]. Differing opinions and unresolved disputes can create a negative environment, hindering collaboration and effective decision-making. In order to resolve this issue, it is crucial to establish open channels of communication where conflicts can be resolved constructively. Promoting team spirit and encouraging collaboration can also help minimize conflict and maintain a positive and productive meeting environment [19], [20].

Table 1 summed up overview of the essential challenges in agile meetings, and possible adequate solutions.

Related previous work is in [21], [22], [23], [24]: analysis and development of a conceptual framework for dealing with agile meetings' problems [21], and of job satisfaction in scrum teams [22], investigation on IT experts' attitude regarding frequency of agile meetings [23]. Additional communication challenges in agile meetings are elaborated in [24].

4. DISCUSSION

Agile meetings often have several challenges that can impede team productivity and satisfaction. The discussion provided here focuses on identifying the main problems frequently arising during agile meetings, as well as on suggesting potential the solutions to these problems. The authors of the paper agree with the relevance of the problems identified and solutions proposed. The authors also suggest several additional/ alternative solutions:

1. Organizing meetings with a clearly defined agenda and goals. This also ensures that the meeting is not delayed and that the meeting is implemented in accordance to the agenda; It may be of importance that the manager is experienced, to pay more attention to these items, because if the team members are dissatisfied with the implementation of the agenda and if the deadline and duration of the meetings is broken, it can lead to many problems and all these can reduce the productivity of the team members;
2. Encouraging the active participation of all team members during the meeting; Team members should support each other by finding the right way to collaborate according to the characters of the team members; The manager must be familiar with his team members and their characters; He or she needs to motivate his or her colleagues to attend meetings actively and work efficiently together;

Table 1. Overview of challenges in agile meetings, and corresponding solutions

Challenge	Sources	Description	Solution
Lack of Focus	[12], [4]	Lack of focus during agile meetings can significantly reduce their effectiveness. Team members can get off topic, wasting valuable time.	Implementing short, well-structured meetings with clearly defined goals and time limits can help maintain focus.
Lack of Engagement	[13], [14]	A lack of team member engagement during meetings can lead to poorer communication and less productive discussions.	Encouraging the active participation of all team members through asking questions and engaging in discussion can significantly improve engagement.
Technological Challenges	[15], [16]	Technical issues such as poor the internet connectivity or inadequate video conferencing tools can hinder agile meetings.	Using reliable access to internet, stable video conferencing tools and providing backup options to deal with technical difficulties can reduce these problems.
Lack of Transparency	[17], [18]	Lack of clarity regarding tasks and goals can lead to confusion and reduced team productivity.	Regular sharing of information on progress and priorities can improve transparency.
Team Conflicts	[19], [20]	Different opinions and conflicts can negatively affect the atmosphere and productivity during meetings.	Establishing open communication channels to resolve conflicts while fostering team spirit and cooperation can help reduce conflicts.



3. Using reliable technological solutions and tools for communication; Software tools need to be made available to employees so they can also work and collaborate from home, thus, employees will be more engaged and motivated to work; and
4. Establishment of open channels for communication within the team to resolve conflicts and eliminate doubts; Team members need to be motivated regularly or frequently, which would reduce the chances of conflicts and dissatisfaction within the team; If team members cannot resolve issues and misunderstandings themselves, the manager needs to react in a good way – as a leader rather than a boss.

The proposed solutions aim to improve the effectiveness of agile meetings. However, limitations to these approaches exist. It is of importance that all these should be considered when analysing the problem.

The limitations to the proposed solutions are as follows:

1. Not all team members may feel comfortable with active participation and initiatives during meetings;
2. Technological problems may still occur despite the use of the available tools;
3. Conflict resolution methods may not be guaranteed to be effective in all cases of team dynamics; and
4. Although the agile approach is more than 20 years old and certified trainings in agile methodologies exist there is still a need to gain better insight regarding its successful implementations in practice and to conduct further research to provide ways to ensure this.

5. CONCLUSION

Agile meetings are one of the crucial activities in contemporary IT companies, driving efficiency, collaboration and adaptability. Identifying, understanding and resolving the issues that arise in agile meetings is key to improving team productivity. To improve collaboration among team members, to help facilitate communication, and create a more efficient work environment several solutions are proposed in the existing literature and here which all can contribute to achieve better overall results.

Suggestions were added to the existing solutions which are valuable, thus expanding the study. They indicate that further investigation is needed to identify more problems and derive corresponding solutions.

By analysing specific meetings in details, teams can effectively implement agile methodologies and improve their project outcomes.

In the future, it is necessary to examine whether there are any additional challenges during meetings of agile teams in IT companies, and to find out whether IT experts would suggest some of their personal solutions to eliminate or avoid these problems. All this can help to increase the satisfaction of employees and customers, boost teams' productivity, and the successful implementation of projects in companies.

In IT companies, where rapid development and adaptability are crucial, successful agile meetings can have a vital role to ensure that projects are delivered efficiently and effectively. Essentially agile meetings are designed to promote effective communication, collaboration and continuous improvement, enabling the agile teams to deliver timely solutions of appropriate values.

REFERENCES

- [1] K. Beck, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries, J. Kern, B. Marick, R. C. Martin, S. Mellor, K. Schwaber, J. Sutherland and D. Thomas, *Manifesto for agile software development*. Wasatch mountains, Utah, USA: Agile Alliance, 2001.[Online]. Available: <https://agilemanifesto.org/>
- [2] A. Cockburn and J. Highsmith, "Agile software development: The people factor," *Computer*, vol. 34, no. 11, pp. 131-133, Nov. 2001. doi: <https://doi.org/10.1109/2.963450>.
- [3] J. Sutherland, *Scrum: The Art of Doing Twice the Work in Half the Time*. 1st ed. New York, New York, USA: Crown Business, 2014.[Online]. Available: https://www.agileleanhouse.com/lib/lib/News/More_Praise_for_Scrum_The_Art_of_Doing_T.pdf.
- [4] K. Schwaber and J. Sutherland, *The Scrum Guide - The Definitive Guide to Scrum: The Rules of the Game*. USA: Ken Schwaber & Jeff Sutherland, Scrumguides.org, 2020.[Online]. Available: <https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf>.
- [5] M. Drury, K. Conboy and K. Power, "Obstacles to decision making in Agile software development teams," *Journal of Systems and Software*, vol. 85, no. 6, pp. 1239-1254, June 2012. doi: <https://doi.org/10.1016/j.jss.2012.01.058>, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0164121212000374>.



- [6] C. de O. Melo, D. S. Cruzes, F. Kon and R. Conradi, "Interpretative case studies on agile team productivity and management," *Information and Software Technology*, vol. 55, no. 2, pp. 412-427, Feb. 2013. doi: <https://doi.org/10.1016/j.infsof.2012.09.004>, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950584912001875>, [Online]. Available: https://www.ime.usp.br/~kon/papers/IST_Melo_2013.pdf
- [7] P. Serrador and J. K. Pinto, "Does Agile work? A quantitative analysis of agile project success," *International Journal of Project Management*, vol. 33, no. 5, pp. 1040-1051, July 2015. doi: <https://doi.org/10.1016/j.ijproman.2015.01.006>, [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0263786315000071>.
- [8] H. Sharp and H. Robinson, "An Ethnographic Study of XP Practice," *Empirical Software Engineering*, vol. 9, pp. 353-375, Dec. 2004. doi: <https://doi.org/10.1023/B:EMSE.0000039884.79385.54>.
- [9] E. Hossain, M. A. Babar and H. -y. Paik, "Using Scrum in Global Software Development: A Systematic Literature Review," in *Proc. 2009 4th IEEE International Conference on Global Software Engineering*, Limerick, Ireland, July 13-16 2009. pp. 175-184, doi: <https://doi.org/10.1109/ICGSE.2009.25>.
- [10] S. Sarker, C. Munson, S. Sarker and S. Chakraborty, "Assessing the Relative Contribution of the Facets of Agility to Distributed ISD Success: An Analytic Hierarchy Process Approach," *European Journal of Information Systems*, vol. 18, no. 4, pp. 285-299, 2009.
- [11] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University Joint Report, Keele and Durham, United Kingdom EBSE Technical Report EBSE-2007-01, 2007. [Online]. Available: <https://cs.ecu.edu/gudivada/research/papers/guidelines-for-se-literature-reviews-summary.pdf>
- [12] H. Kniberg and M. Skarin, *Kanban and Scrum – Making the Most of Both*. USA: C4Media, Publisher of InfoQ.com, 2010. [Online]. Available: <https://www.infoq.com/minibooks/kanban-scrum-mini-book/>, [Online]. Available: <https://www.agileleanhouse.com/lib/lib/People/HenrikKniberg/KanbanAndScrumInfoQVersionFINAL.pdf>
- [13] E. Gottesdiener, *Requirements by Collaboration: Workshops for Defining Needs*, Boston, Massachusetts, USA: AddisonWesley, 2002.
- [14] E. Derby and D. Larsen, *Agile Retrospectives: Making Good Teams Great*, Raleigh, North Carolina and Dallas, Texas, USA: Pragmatic Bookshelf, 2006. [Online]. Available: <https://agile.2ia.net/Agile%20Retrospectives.pdf>
- [15] A. Stellman and J. Greene, *Learning Agile: Understanding Scrum, XP, Lean, and Kanban*, Sebastopol, California, USA: O'Reilly Media, Inc., 2014. [Online]. Available: <https://ecole-management.com/BOOKS/2.pdf>
- [16] J. Bablo, B. Marcinkowski and A. Przybylek, "Overcoming Challenges of Virtual Scrum Teams: Lessons Learned Through an Action Research Study," In: Stettina, C. J., Garbajosa, J., Kruchten, P. (eds) *Agile Processes in Software Engineering and Extreme Programming*, XP 2023, Lecture Notes in Business Information Processing - LNBIIP, Cham, Switzerland: Springer, vol. 475, pp. 34-49, 2023. doi: https://doi.org/10.1007/978-3-031-33976-9_3
- [17] C. Larman, *Agile and Iterative Development: A Manager's Guide*, Boston, Massachusetts, USA: Addison-Wesley, 2004.
- [18] K. S. Rubin, *Essential Scrum: A Practical Guide to the Most Popular Agile Process*, Upper Saddle River, NJ, Boston, Indianapolis, San Francisco, New York, Toronto, Montreal, London, Munich, Paris, Madrid, Capetown, Sydney, Tokyo, Singapore, Mexico City, USA & International ed. Addison-Wesley & Pearson Education, Inc., 2012.
- [19] L. Adkins, *Coaching Agile Teams: A Companion for ScrumMasters, Agile Coaches, and Project Managers in Transition*, Upper Saddle River, NJ, Boston, Indianapolis, San Francisco, New York, Toronto, Montreal, London, Munich, Paris, Madrid, Capetown, Sydney, Tokyo, Singapore, Mexico City, USA & International ed.: Addison-Wesley & Pearson Education, Inc., 2010.
- [20] T. DeMarco and T. Lister, *Peopleware: Productive Projects and Teams*, 3rd ed., Upper Saddle River, NJ, Boston, Indianapolis, San Francisco, New York, Toronto, Montreal, London, Munich, Paris, Madrid, Capetown, Sydney, Tokyo, Singapore, Mexico City, USA & International ed.: Addison-Wesley & Pearson Education, Inc., 2013.
- [21] M. Gaborov, Ž. Stojanov and S. Popov, "Thematic analysis and development of a conceptual framework for agile meetings' problems," *Journal of Engineering Management and Competitiveness - JEMC*, vol. 13, no. 1, pp. 61-70, June 2023. doi: <https://doi.org/10.5937/JEMC2301061G> [Online]. Available: <https://scindeks-clanci.ceon.rs/data/pdf/2334-9638/2023/2334-96382301061G.pdf>
- [22] M. Gaborov, D. Kovač, N. Tasić, D. Kreculj and N. Ratković Kovačević, "Job Satisfaction in Scrum Teams," in: *Proceedings of the 5th Virtual International Conference on Path to a Knowledge Society- Managing Risks and Innovation - PaKSoM 2023*, On-line & Niš, Serbia, Oct. 23-24 2023, pp. 335-339.
- [23] M. Gaborov, Z. Stojanov, S. Popov, D. Kovac, I. Vecstajn, G. Jotanovic and G. Jausevac, "Investigating IT Experts' Attitudes Toward the Frequency of Agile Meetings: A Thematic Analysis," in: *Proceedings of the 14th International Conference on Applied Internet and Information Technologies - AIIT 2024*, Zrenjanin, Serbia, Nov. 8 2024, pp. 200-207. [Online]. Available: <http://www.tfzr.rs/aiit/files/AIIT2024%20Proceedings.pdf>
- [24] M. Gaborov, M. Kovačević, N. Tasić, D. Kreculj, N. Ratković Kovačević and D. Kovač, "Communication Challenges in Agile Meetings," in: *Proceedings of the 14th International Conference on Applied Internet and Information Technologies - AIIT 2024*, Zrenjanin, Serbia, Nov. 8 2024, pp. 334-340. [Online]. Available: <http://www.tfzr.rs/aiit/files/AIIT2024%20Proceedings.pdf>



DIGITAL WATERMARKING IN IMAGES USING DCT: EMBEDDING, EXTRACTION AND VISUAL QUALITY PRESERVATION

Nenad Stanojević^{1*},
[0009-0004-4158-6412]

Petar Spalević¹,
[0000-0002-6867-7259]

Stefan Panić²,
[0000-0002-5868-1764]

Đorđe Šarčević³,
[0000-0003-0746-744X]

Srđan Mitrović⁴
[0009-0006-5105-4270]

¹University of Priština,
Faculty of Technical Sciences,
Kosovska Mitrovica, Serbia

²University of Priština,
Faculty of Sciences,
Kosovska Mitrovica, Serbia

³Department of Medical and
Business-Technological Studies,
Academy of Professional Studies Šabac,
Šabac, Serbia

⁴Singidunum University
Belgrade, Serbia

Correspondence:

Nenad Stanojević

e-mail:

nenads25@gmail.com

Abstract:

This paper presents an algorithm for embedding and extracting a digital watermark in an image using the Discrete Cosine Transform (DCT). The standard Lenna image was used as the test image, while the watermark was implemented in the form of a chessboard, defined within a simple matrix structure to facilitate implementation. The simulation of the process was carried out using MATLAB software support. This approach enables the visualization of the process, significantly simplifying the embedding and extraction of the digital watermark. In addition to its application in digital image protection, this model can serve as a foundation for analyzing the impact of various transmission channels on an image, such as atmospheric channels in wireless communications. Beyond the DCT, the proposed approach can also be adapted to other frequency domain transformations, such as the Discrete Fourier Transform (DFT) or the Discrete Wavelet Transform (DWT).

Keywords:

Discrete Cosine Transform, Digital Watermark, Algorithm, Frequency Domain.

INTRODUCTION

One of the most challenging issues in today's digital era is protecting data from unauthorized access, sharing, and duplication. The modern digital environment allows easy access to documents while sharing images and video content via social media has become particularly popular. This raises the question of how to protect digital content from misuse. One possible solution is embedding a digital watermark into a selected digital asset to verify authenticity. A digital watermark is a process of embedding protective information into a digital asset, allowing this information to be later extracted for authenticity verification. Digital watermarks can be either visible or invisible. A visible watermark is immediately noticeable within the digital content, whereas detecting an invisible watermark requires appropriate computing tools and software. This technique serves a dual purpose: on one hand, it enables the identification of the source and author of the digital content, while on the other, it provides authors with protection against unauthorized use [1], [2].



Various studies have been conducted to find the best approach for designing a model for embedding and extracting digital watermarks. It has been observed that after watermark extraction, distortions occur both in the image itself and in the watermark. Studies [3], [4] present different embedding and extraction models that are structurally most compatible with the human visual system.

The most suitable domains for digital watermarking are the frequency domains. To achieve imperceptibility, the watermark should be embedded in the high-frequency components of the transmitted signal. Conversely, for a robust digital watermark, embedding should be restricted to low-frequency components, with the low-frequency component serving as the 'host' for watermark insertion. In the paper [5], an algorithm combining different transformation schemes (DFT, DCT, SVD, and DWT) is presented. It has been proven that by combining different models, the resistance, robustness, and imperceptibility of the embedded digital watermarks are enhanced. However, due to their complexity, the implementation of such models can often be financially unfeasible. Paper [6] proposes a digital watermarking system for color images in the spatial domain, combining the strengths of both spatial and frequency-domain schemes. The system uses different quantization steps for embedding and blind extraction of the watermark, improving invisibility and reducing the pixel modification range. Experimental results demonstrate that this system offers better invisibility, robustness, and shorter execution time compared to existing methods.

Various methods have been considered with the aim of implementing a digital watermark that satisfies both the conditions of imperceptibility and robustness, while also addressing financial aspects. In [7], a scheme for removing visible watermarks with the possibility of reversible image recovery is presented. This model is one of the most efficient, as it allows for the complete reconstruction of the original image. However, a certain level of degradation is unavoidable.

This paper analyzes the process of embedding and extracting a digital watermark using the Discrete Cosine Transform (DCT), which belongs to the class of frequency domain transformations. The MATLAB environment was used for process analysis, providing a simplified representation of the embedding and extraction procedures for digital watermarking.

2. RELATED WORK

In this section, we examine modern techniques for embedding digital watermarks, focusing on frequency-domain-based methods. We analyze the efficiency and limitations of existing techniques to identify gaps in this key research field.

In [8], [9], a two-dimensional discrete cosine transform (2D-DCT) model for embedding digital watermarks is presented. This method uses mid-frequency components for watermark embedding, ensuring its imperceptibility. The model employs color images for both the original image and the watermark, providing high resistance to attacks, minimal visibility of the embedded watermark, and enabling a high embedding capacity. This approach is particularly suitable for embedding company color logos, contributing to copyright protection. To enhance security and robustness, the color watermark is divided into its primary components (red, blue, and green), followed by Arnold transformation to permute pixels using a private key. Experimental analysis has demonstrated high resistance to compression, filtering, cropping, and scaling. Peak signal-to-noise ratio (PSNR) values exceeded 36 dB, indicating high imperceptibility of the embedded watermark.

In [10], a robust and secure system for embedding digital watermarks based on the encryption of a random binary sequence is presented. This model utilizes second-level DWT (2DWT) combined with DCT to enhance watermark protection and reduce image distortion. DWT is used to decompose the image and isolate low-frequency sub-bands (LL), while 2DWT is applied to improve resistance to filtering and reduce visual artifacts. The use of DCT further enhances watermark embedding by modifying AC coefficients without significantly degrading image quality. Experimental analysis has demonstrated the high robustness of the method against various image processing attacks, including scaling, filtering, compression, sharpening, and noise addition. The advantage of this approach lies in its low error rate in watermark reconstruction along with high imperceptibility, proving its suitability for copyright protection and cybersecurity applications. R. Soundrapandian et al. in [11] proposed a model that also employs the DWT-DCT transformation but applies different protection strategies. The model divides the image into regions of interest (ROI) and non-interest regions (NIR), embedding watermarks by quantizing low-frequency coefficients and using zigzag sorting for DCT coefficients. This model has been particularly useful in securing clinical medical images, where watermarks are



used to store sensitive patient data. The system not only ensures the confidentiality and integrity of the medical data but also exhibits high robustness, with PSNR values exceeding 40 dB after watermark extraction. The method has shown resilience against common image processing attacks, ensuring the security of medical images in applications such as telemedicine and patient data protection.

H. Cao et al. [12] present a model for embedding a robust reversible digital watermark using DFT in the spatial domain. The proposed approach utilizes the relationship between the DC component of the DFT and pixel values in the spatial domain to avoid explicit DFT and IDFT transformations, significantly reducing execution time compared to traditional DFT-based watermarking methods. The quantization step is optimized to achieve a balance between imperceptibility and robustness. Robustness ensures copyright protection, while reversibility preserves image integrity and authentication. Experimental analysis has shown that the model enables watermark extraction without loss of original image quality, with good robustness, imperceptibility, and high embedding capacity. Key concepts of this approach include two-dimensional discrete Fourier transform (2D-DFT), DC component modification, and encryption based on logistic chaotic mapping. However, the model's resistance to noise attacks and pixel cropping requires further optimization.

The presented models demonstrate different approaches to digital watermark embedding. Most of the examined approaches achieve good resistance to attacks, high embedding capacity, and reliability; however, they share a common characteristic - focusing solely on invisible watermarks. Invisible watermarks are desirable in many scenarios, particularly in copyright protection, forensic analysis, and digital content authentication, where the goal is to preserve image quality without noticeable alterations. However, in certain cases, there is a need for visible watermarks, such as preventing unauthorized use of images online, marking ownership of digital content, or branding media materials. Visible watermarks enable direct ownership identification and are frequently used in news agencies, stock photo databases, and video platforms, where it is crucial to emphasize the source of an image or video.

While invisible watermarks are effective for hidden content protection, current methods lack flexibility in choosing between visible and invisible watermarking. The development of adaptive techniques that allow controlled watermark visibility could represent a significant advancement in this field. Integrating existing

frequency-domain methods with adaptive techniques could enable dynamic selection of watermark visibility depending on the application.

3. DCT (DISCRETE COSINE TRANSFORM)

DCT is one of the most commonly used transformations in the frequency domain. This transformation operates only with real values and decomposes a signal into a combination of cosine functions of different frequencies.

If an image in the spatial domain is represented as the function $f(x, y)$, where it is defined in two dimensions (x and y), the Discrete Cosine Transform (DCT) projects this image into the frequency domain, where the transformed function is denoted as $F(u, v)$.

$$f(x, y) \xrightarrow{DCT} F(u, v)$$

Equation 1. Transformation from the spatial domain to the frequency domain using DCT

If observation is performed in a one-dimensional domain, the transformation can be represented by the expression:

$$F(u) = c \cdot f(x)$$

Equation 2. One – dimensional DCT

Where c is the cosine transformation matrix. When the matrix c is multiplied by the function $f(x)$, the entire expression takes the form:

$$F(u) = a(u) \sum_x^{n-1} f(x) \cos \left[\frac{(2x+1)\pi u}{2n} \right] \text{ for } 0 \leq u \leq n-1$$

Equation 3. DCT - transform expression

$$\text{where: } \begin{cases} a(u) = \sqrt{\frac{1}{n}} & \text{if } u = 0 \\ a(u) = \sqrt{\frac{2}{n}} & \text{if } u \neq 0 \end{cases}$$

The previous expression represents the matrix used for the 1-D DCT transformation. The process of the 2-D DCT transformation is implemented using the expression:

$$F(u, v) = c \cdot f(x, y) \cdot c^T$$

Equation 4. Two – dimensional DCT

Where c^T represents the transposed matrix, and the expression for the 2-D DCT transformation is [4], [13]:

$$F(u, v) = a(u)a(v) \sum_x^{n-1} \sum_y^{m-1} f(x, y) \cos \left[\frac{(2x+1)\pi u}{2n} \right] \cos \left[\frac{(2y+1)\pi v}{2m} \right]$$

Equation 5. Two – dimensional DCT – transform expression



$$\text{where: } \begin{cases} a(u) = \sqrt{\frac{1}{n}} & \text{if } u = 0 \\ a(u) = \sqrt{\frac{2}{n}} & \text{if } u \neq 0 \end{cases}$$

$$\text{where: } \begin{cases} a(v) = \sqrt{\frac{1}{m}} & \text{if } v = 0 \\ a(v) = \sqrt{\frac{2}{m}} & \text{if } v \neq 0 \end{cases}$$

If the image is divided into blocks, and each of these blocks is made up of pixels, each pixel represents a small part of the image. The DCT transformation allows these pixels to be represented through a cosine function, and what the previous expression enables is the calculation of coefficients that carry information about the details of the image [13], [14].

The process of embedding a digital watermark into an image using the DCT

For the implementation of this procedure, MATLAB programming support was used. The process of embedding a digital watermark into the image was carried out using the appropriate algorithm.

Step 1: Loading the original image (Figure 1).

Step 2: In this step, the image is vectorially separated into its basic RGB components. Then, using DCT, each component is individually transformed into the frequency domain. The DCT transformation allows the pixels to be represented through a cosine function, and equation (5) enables the calculation of coefficients that carry information about the details of the image [14], [15], [16], [17]. The selected image of dimensions $M \times N$ is decomposed into non-overlapping blocks of size $m \times n$, and then each block is transformed into the corresponding DCT coefficients from equation (5).

The blocks resulting from the DCT transformation can be divided into three distinct frequency ranges: low, medium, and high. The low-frequency range contains the most significant information about the image, and modifying it leads to perceptual changes in the image. On the other hand, the high-frequency components have less importance and can be removed for compression purposes [15], [18]. When determining the coefficients, each value is divided according to its weighting factor. Quantization groups the weighting factors in such a way that it maximizes the number of components close to zero while maintaining the image quality. After quantization, a zig-zag transformation is applied, grouping identical values [18], [19], [20], [21], [22]. Figure 2 shows the arrangement of coefficients and their graphical representation.

Step 3: A digital watermark in the form of a chessboard is created by appropriately arranging binary codes "1" and "0" in a matrix structure. The watermark A is created with a resolution of 256×256 ($a \times b$). The matrix is created based on the principle of arranging the fields of a chessboard, where each field consists of 32×32 binary codes, with "1" representing white squares and "0" representing black squares (Figure 3).

Step 4: In this step, the digital watermark is embedded. The embedding process is straightforward, where the binary components in the matrix of the digital watermark are added to the selected DCT coefficients, thus modifying the coefficient values.

$$F'(u,v) = \begin{cases} F(u,v) + Ku \cdot A(a,b) & \text{for } 0 \leq u < a \text{ and } 0 \leq v < b \\ F(u,v) & \text{For all other values} \end{cases}$$

Equation 6. Expression for digital watermark embedding

Where $F'(u,v)$ is the image with the binary digital watermark in the frequency domain. The selection of coefficients is done by adding the first 256×256 elements of the DCT image (for each component individually) to the elements of matrix A. During the addition of the elements of the matrix, the embedding coefficient Ku is also defined, which represents the level of transparency of the digital watermark.

Step 5: After applying the inverse DCT (iDCT), the image with the embedded digital watermark is obtained (Figure 4). The embedding coefficient Ku plays a crucial role in the implementation of the watermark - at certain minimal values, the digital watermark remains imperceptible, while at certain maximal values, it can significantly degrade the image quality, making it unusable until the watermark extraction is performed.

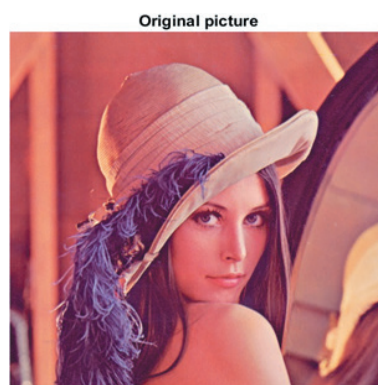


Figure 1. Original picture



Figure 2. DCT zig-zag distribution of coefficients and frequency distribution of coefficients, respectively [23]

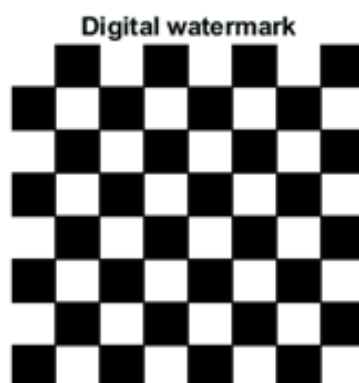


Figure 3. Digital watermark chessboard

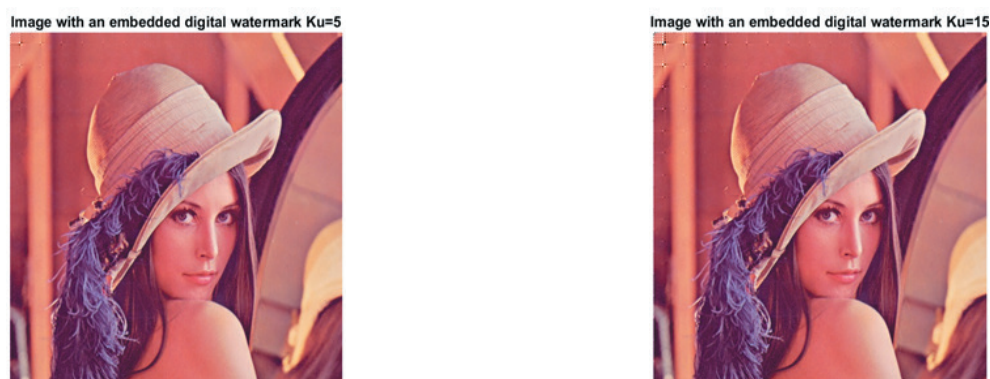


Figure 4. Images with the embedded digital watermark for $Ku=5$ and $Ku=15$, respectively



5. ALGORITHM FOR EXTRACTING THE WATERMARK FROM THE IMAGE

When extracting a watermark from an image, three methods are typically used [16], [19]:

1. Blind method – this approach does not require the original image for watermark extraction.
2. Non-blind method – the original image is necessary for extracting the watermark.
3. Semi-blind method – this method utilizes the original digital watermark or other relevant information for detection.

According to the method of watermark extraction, the Semi-blind method is used in this paper. The extraction process consists of the following steps:

Step 1: The process of determining the coefficients in the frequency domain using DCT is repeated, but this time it is performed on the image with the embedded digital watermark.

Step 2: From the image composed of coefficients created by DCT, the matrix of the digital watermark is subtracted with the same transparency coefficient as in the implementation.

$$F(u,v) = \begin{cases} F'(u,v) - Ku \cdot A(a,b) & \text{for } 0 \leq u < a \text{ and } 0 \leq v < b \\ F'(u,v) & \text{For all other values} \end{cases}$$

Equation 7. Expression for digital watermark extraction

Step 3: After removing the watermark, the final step involves applying the iDCT, which returns the coefficients to their original state, as before the transformation (Figure 5).

The image obtained after extraction demonstrates exceptional quality, with no visible signs of damage that could have resulted from the application of the DCT transformation and the process of embedding the digital watermark.

However, the proposed model has a certain drawback in terms of security. In addition to the simple scheme of the watermark itself, an additional challenge is the fact that the digital watermark is embedded in the lower frequencies of the DCT image, which can reduce its resistance to attacks. The level of security can be significantly improved by embedding the digital watermark into the middle or higher frequency components of the DCT transformation. Moreover, adding an encryption element, such as AES (Advanced Encryption Standard), RSA (Rivest-Shamir-Adleman), or chaos-based encryption methods, could further strengthen the protection and eliminate this drawback. The effectiveness of these methods can be further enhanced by combining them with appropriate transformation domains, such as DWT (Discrete Wavelet Transform) or SVD (Singular Value Decomposition), achieving greater resistance to various attacks and signal degradation.

These possibilities provide a potential foundation for further development of the model and future research in this area.

Image after digital watermark removal



Figure 5. Code for iDCT and the image after watermark extraction



6. CONCLUSION

This paper explores the process of embedding and extracting a digital watermark in images using the Discrete Cosine Transform (DCT). A method is presented that enables watermark embedding while preserving high image quality. The extraction is performed using the semi-blind method, without requiring the original image, while the inverse DCT (iDCT) is used to restore the image to its original state.

The results show that the proposed method does not degrade image quality after watermark extraction. Additionally, by increasing the embedding coefficient K_u , the transparency level of the digital watermark can be adjusted. Due to its simplicity and ability to preserve visual quality, this approach can facilitate image transmission through communication channels.

The proposed method has potential applications in copyright protection and digital content authentication. Future research could focus on enhancing its robustness against compression and extending its application to dynamic media, such as video.

REFERENCES

- [1] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich and T. Kalker, "Digital Watermarking and Steganography," *Morgan Kaufman Publisher*, 2008.
- [2] O. Evsutin, A. Melman and R. Meshcheryakov, "Digital steganography and watermarking for digital images: A review of current research directions," *IEEE Access*, vol. 8, pp. 166589-166611, 2020.
- [3] A. Dixit and R. Dixit, "A Review on Digital Image Watermarking Techniques," *International Journal of Image, Graphics and Signal Processing*, pp. 56-66, 2017.
- [4] M. Kaur, S. Jindal and S. Behal, "A study of digital image watermarking," *IJREAS*, vol. 2, p. 126-136, 2012.
- [5] J. Varghese, B. O. Hussain, A. Razak T. and S. Sabash, "A Hybrid Digital Image Watermarking Scheme Incorporating DWT, DFT, DCT and SVD Transformations," *PeerJ Computer Science*, vol. 10, no. 1A, pp. 113-130, 2022.
- [6] Z. Yuan, O. Su, D. Liu and X. Zhang, "A blind image watermarking scheme combining spatial domain and frequency domain," *The visual computer*, vol. 37, p. 1867-1881, 2021.
- [7] C. C. Chen, Y. H. Tsai and H. C. Yeh, "Difference-expansion based reversible and visible image watermarking scheme," *Multimedia Tools Appl.*, vol. 76, p. 8497-8516, 2017.
- [8] Z. Yuan, D. Liu, X. Zhang and Q. Su, "New image blind watermarking method based on two-dimensional discrete cosine transform," *Optik*, 2020.
- [9] H. Wang, Z. Yuan, S. Chen and Q. Su, "Embedding color watermark image to color host image based on 2D-DCT," *Optik*, 2023.
- [10] N. Hasan, M. S. Islam, W. Chen, M. A. Kabir and S. Al-Ahmadi, "Encryption Based Image Watermarking Algorithm in 2DWT-DCT Domains," *Sensors*, vol. 21, no. 16, 2021.
- [11] R. Soundrapandian, K. Rajendiran, A. Gurunathan, A. Victor and R. Selvanambi, "Analysis of DWT-DCT watermarking algorithm on digital medical imaging," *Journal of Medical Imaging*, vol. 11, no. 1, 2024.
- [12] H. Cao, H. Fu, Y. Sun, S. Chen and Q. Su, "Robust and reversible color image watermarking based on DFT in the spatial domain," *Optik*, 2022.
- [13] A. B. Watson and NASA Ames Research Center, "Image compression using discrete cosine transform," *Mathematica journal*, vol. 4, no. 1, pp. 81-88, 1994.
- [14] V. Singh, "Digital Watermarking: A Tutorial," *JSAT*, 2011.
- [15] S. Roy and A. K. Pal, "A blind DCT based color watermarking algorithm for embedding multiple watermarks," *AEU-International Journal of Electronics and Communications*, vol. 72, pp. 149-161, 2017.
- [16] N. F. Johnson and S. Katzenbeisser, "A survey of steganographic techniques," In *Information hiding*, pp. 43-78, 2000.
- [17] W. Chu, "DCT-based image watermarking using subsampling," *IEEE Trans. Multimedia*, vol. 5, no. 1, p. 34 - 38, 2003.
- [18] M. Jiansheng, L. Sukang and T. Xiaomei, "A digital watermarking algorithm based on DCT and DWT," In *Proceedings. The 2009 International Symposium on Web Information Systems and Applications*, (WISA 2009), p. 104, 2009.
- [19] M. I. Khan, M. M. Rahman and M. I. Sarker, "Digital watermarking for image authentication based on combined dct, dwt and svd transformation," *arXiv preprint arXiv:1307.6328*, 2013.
- [20] R. A. Asmara, and R. Agustina, "Comparison of discrete cosine transforms (DCT), discrete Fourier transforms (DFT), and discrete wavelet transforms (DWT) in digital image watermarking," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 2, 2017.
- [21] S. Katzenbeisser and F. Petitcolas, "Information hiding," *Artech house*, 2016.
- [22] S. M. Aghajanzadeh and M. Uysal, "Diversity-multiplexing trade-off in coherent free-space optical systems with multiple receivers," *J. Opt. Commun. Netw*, vol. 2, pp. 1087-1094, 2010.
- [23] M. Baechler and J. L. Bloechle, "Labeled images verification using gaussian mixture models," In *Proceedings of the 2009 ACM symposium on Applied Computing*, pp. 1331-1335, 2009.



PERFORMANCE OPTIMIZATION OF FILE SYSTEMS FOR DOCKER CONTAINERS

Sava Stanišić^{1*},
[0009-0002-3118-0537]

Borislav Đorđević²,
[0000-0002-6145-4490]

Olga Ristić^{1,3},
[0000-0002-1723-0940]

Ivan Tot³
[0000-0002-5862-9042]

¹Faculty of Technical Sciences,
Čačak Serbia

²Mihajlo Pupin Institute,
Belgrade, Serbia

³Military Academy,
Belgrade, Serbia

Abstract:

The performance of file systems plays a crucial role in containerized environments, directly affecting the efficiency and scalability of applications deployed using Docker. This paper explores the impact of various file systems on Docker container performance, focusing on metrics such as I/O throughput, latency, and resource usage. Through an experimental evaluation of file systems, including OverlayFS, Advanced Multi-Layered Unification File System (AUFS), and B-Tree File System (Btrfs), their behavior under different workloads is analyzed. Additionally, the techniques to improve file system performance are proposed, leveraging DevOps tools for monitoring and automation. The findings of this research offer actionable insights for system administrators and DevOps engineers seeking to optimize container storage performance in both cloud and on-premises environments.

Keywords:

Docker, File Systems, Performance Optimization, Container Storage.

INTRODUCTION

Containerization has emerged as a fundamental technology in modern software development, enabling efficient resource utilization, application portability, and scalable deployment models. Docker, one of the most widely adopted containerization platforms, has become a critical tool for cloud-native development and DevOps practices. However, the performance of containerized applications is significantly influenced by the underlying storage architecture, particularly the file system.

File systems play a vital role in managing data storage and retrieval operations within Docker environments. These systems handle complex storage structures by layering data and maintaining file consistency across container instances. Commonly used file systems such as OverlayFS, Advanced Multi-Layered Unification File System (AUFS), and B-Tree File System (Btrfs) are essential for Docker's storage capabilities. Each file system exhibits distinct characteristics and performance implications based on its design and operational principles.

Correspondence:

Sava Stanišić

e-mail:

sava.stanasic@vs.rs





In high-performance computing and large-scale deployments, optimizing file system performance becomes critical. The efficiency of input/output (I/O) operations, data caching mechanisms, and resource utilization can determine the responsiveness and stability of containerized applications. Understanding the impact of different file system configurations on these parameters is crucial for achieving optimal performance.

This research investigates the performance characteristics of various file systems used in Docker environments, with a focus on identifying configurations that enhance storage efficiency and system responsiveness. I/O throughput, latency, and resource usage are analyzed across various workloads. Based on the experimental findings, recommendations for selecting and configuring file systems for diverse application types are proposed.

The insights presented in this study contribute to the optimization of container-based deployments in both cloud and on-premises environments, providing valuable guidance for system administrators and DevOps engineers.

2. BACKGROUND AND RELATED WORK

The efficient management of file systems in containerized environments has garnered significant attention due to the increasing adoption of containerization technologies. Docker, as a leading platform in this domain, supports multiple file systems designed to manage data storage, retrieval, and consistency across container layers. This chapter provides an overview of Docker's storage architecture, the characteristics of various supported file systems, and a review of existing research on file system performance in containerized environments.

2.1. PREVIOUS RESEARCH

The performance and efficiency of containerized workloads are heavily influenced by underlying file system architectures and storage drivers. Early work by Felter et al. [1] provided a foundational analysis of Linux container performance, highlighting the critical role of storage driver selection in I/O-intensive applications. Building on this, Ferreira et al. [2] conducted a comparative study of Docker storage drivers, demonstrating that OverlayFS achieves superior read/write throughput for web applications, on the other hand, Btrfs excels in scenarios requiring frequent large-scale dataset modifications.

In cloud-native environments, Tarasov et al. [3] evaluated OverlayFS optimizations, showing that its copy-on-write mechanism reduces container startup latency by up to 40% compared to traditional union file systems. This aligns with findings by Cilic et al. [4], who demonstrated that OverlayFS minimizes disk I/O overhead in clusters by leveraging page cache sharing across container layers.

2.2. RESEARCH HYPOTHESIS AND QUESTIONS

Hypothesis: Optimizing file system selection and configuration enhances Docker container performance in various workload scenarios.

Research Questions:

1. How do different file systems affect Docker container performance under various workloads?
2. What techniques can improve the efficiency of file systems in containerized environments?
3. How does file system choice impact resource utilization in high-performance computing scenarios?

2.3. DOCKER STORAGE ARCHITECTURE

Docker's storage system is designed to provide scalable and efficient data management for containerized applications. It employs a layered architecture where file systems play a crucial role in storing and managing data. Each container in Docker is built on top of a read-only image layer, with writable layers on top to capture changes made during container execution.

Union file systems such as OverlayFS and AUFS are commonly used to implement this layered architecture. These file systems enable efficient data storage by merging multiple file system layers into a unified view. Btrfs, a copy-on-write (CoW) file system, offers advanced features such as snapshots and dynamic disk allocation, making it suitable for complex storage requirements.

2.4. CHARACTERISTICS OF COMMON DOCKER FILE SYSTEMS

OverlayFS: A modern union file system designed for performance and efficiency. OverlayFS merges multiple directories into a single unified view and is optimized for Docker's layered architecture. Its simplicity and high performance have made it the default file system for Docker on many Linux distributions.

AUFS (Advanced Multi-Layered Unification File System): One of the earliest union file systems used by Docker. While still supported, it has been largely replaced by OverlayFS due to better performance and kernel support.



Btrfs: A CoW file system known for its advanced features, including snapshots, subvolumes, and dynamic disk space allocation. Btrfs offers high scalability and flexibility but may introduce additional resource overhead compared to other file systems.

3. EVALUATION AND RESULTS

The performance evaluation results of file systems in the Docker container are presented here. The focus was on three widely used file systems—OverlayFS, AUFS, and Btrfs—and their performance was analyzed under diverse workloads, including database operations, web server I/O, and machine learning tasks. Key metrics such as I/O throughput, latency, and resource utilization are measured and compared to determine the most suitable file system for specific use cases.

3.1. WORKLOAD SPECIFIC PERFORMANCE

The performance of OverlayFS, AUFS, and Btrfs was evaluated under three distinct workloads: database operations, web server I/O, and machine learning training. The results are summarized below.

3.1.1. Database Workload

Setup: Simulated a MySQL database with 10,000 transactions, representing a write-intensive workload.

Results:

- Throughput: Btrfs achieved the highest throughput (1,500 IOPS), followed by OverlayFS (1,200 IOPS) and AUFS (1,000 IOPS).
- Latency: OverlayFS had the lowest average latency (2.8 ms), while Btrfs and AUFS averaged 3.5 ms and 4.0 ms, respectively.
- CPU Usage: Btrfs consumed 25% more CPU than OverlayFS and AUFS due to its advanced features like CoW and snapshots.

Analysis: Btrfs's high throughput is attributed to its efficient handling of concurrent writes, and this comes at the cost of increased CPU usage. OverlayFS, on the other hand, provides a good balance of performance and resource efficiency for database workloads.

3.1.2. Web Server Workload

Setup: Simulated an Nginx web server serving 10,000 small files, representing a read-intensive workload.

Results:

- Throughput: OverlayFS achieved the highest throughput (900 IOPS), outperforming AUFS (800 IOPS) and Btrfs (750 IOPS).

- Latency: OverlayFS had the lowest average latency (1.5 ms), while AUFS and Btrfs averaged 2.0 ms and 2.5 ms, respectively.
- Memory Usage: OverlayFS used 10% less memory than AUFS and Btrfs.

Analysis: OverlayFS's efficient merging mechanism and lightweight design make it ideal for read-heavy workloads like web servers. AUFS, while still performant, lags due to its older architecture.

3.1.3. Machine Learning Workload

Setup: Simulated a TensorFlow training job with large sequential reads and writes, representing a data-intensive workload.

Results:

- Throughput: Btrfs achieved the highest throughput (600 MB/s), followed by OverlayFS (500 MB/s) and AUFS (450 MB/s).
- Latency: Btrfs had the lowest latency (3.8 ms), while OverlayFS and AUFS averaged 4.5 ms and 5.0 ms, respectively.
- Disk Usage: Btrfs consumed 20% more disk space due to its copy-on-write and compression features.

Analysis: Btrfs's advanced features, such as dynamic disk allocation and compression, make it well-suited for data-intensive workloads like machine learning. However, its higher resource consumption may be a limiting factor in resource-constrained environments.

3.2. FILE SYSTEM COMPARISON

The results are summarized in Table 1.

The graphical representation of the results is given in the Figure 1.

3.3. IMPACT OF DOCKER STORAGE DRIVERS

The performance of Docker storage drivers (overlay2 and aufs) was also evaluated with each file system [5][6]. The results are summarized below.

- overlay2: Consistently performed well across all file systems, with minimal overhead.
- aufs: Showed higher latency for write-intensive workloads, particularly with Btrfs and AUFS.



Table 1. File System Performance Summary

File System	Throughput (IOPS)	Latency (ms)	CPU Usage (%)	Memory Usage (%)
OverlayFS	1200	2.8	20	15
AUFS	1000	4.0	22	18
Btrfs	1500	3.5	25	20

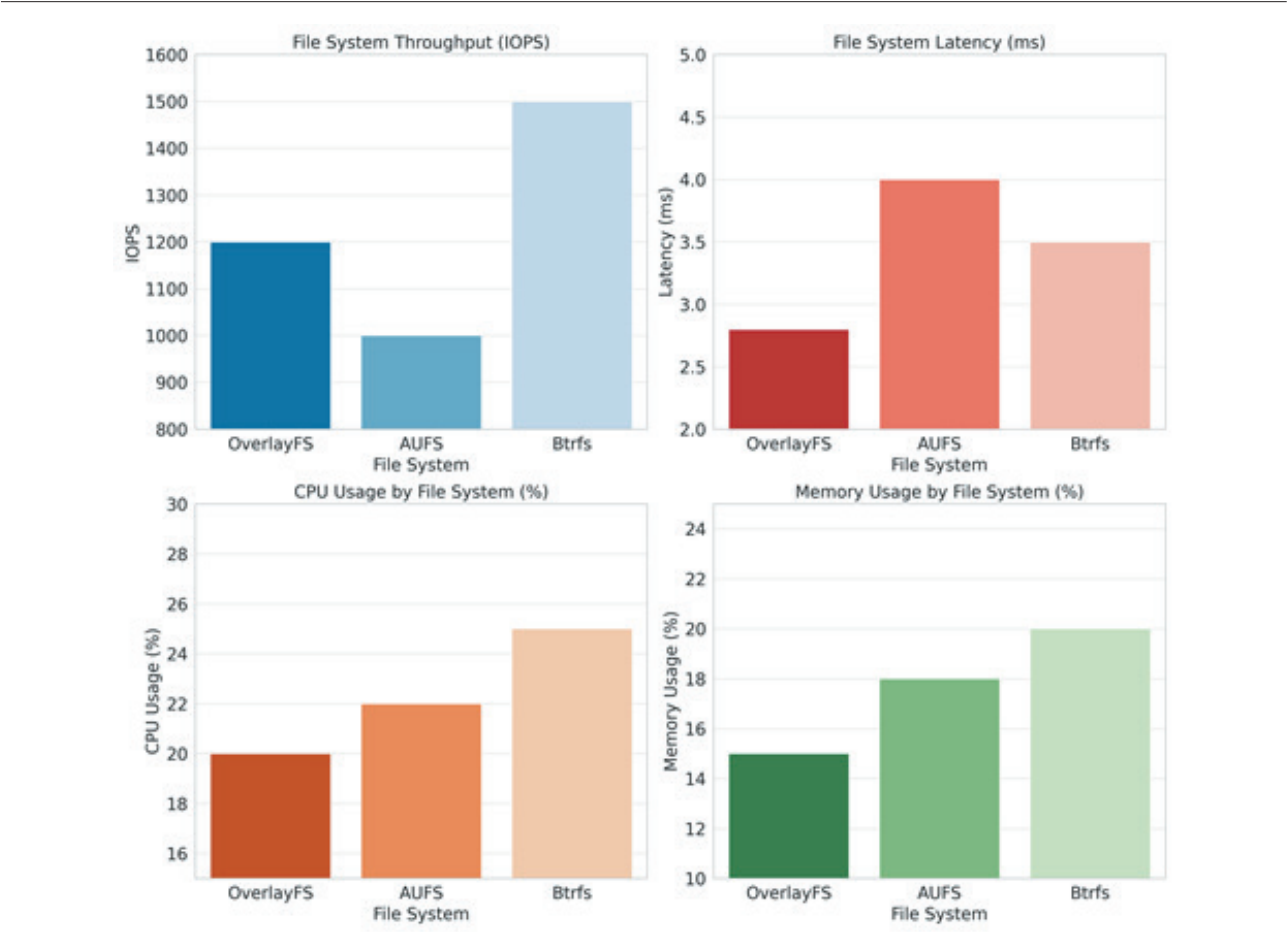


Figure 1. The graphical representation of the results

4. BENCHMARKING TOOLS AND METHODOLOGY

To evaluate the performance of OverlayFS, AUFS, and Btrfs in Docker environments [7], a combination of industry-standard benchmarking tools and custom scripts was employed. The process involved setting up a controlled test environment, defining workloads, and measuring key performance metrics. Below, the tools, setup, and methodology are described in detail.

4.1. BENCHMARKING TOOLS

The following tools were used to measure file system performance:

- Fio:
 - A versatile tool for benchmarking I/O performance.
 - Supports a wide range of I/O patterns (e.g., sequential, random, read, write)
 - Used to measure throughput (IOPS), latency, and bandwidth.
- Sysbench:
 - A modular, cross-platform benchmarking tool.
 - Used for database workload simulations (e.g., MySQL transactions).



- Measures transaction throughput, latency, and resource usage.
- Docker Stats:
 - A built-in Docker tool for monitoring container resource usage.
 - Used to measure CPU, memory, and disk I/O utilization during tests.
- Bonnie++:
 - A benchmark tool for testing file system performance.
 - Used to evaluate sequential and random I/O performance.
- Custom Scripts:
 - Bash scripts were developed to automate workload execution, data collection, and result analysis.
 - These scripts ensured consistency across multiple test runs.

The scripts provided in Appendix A collectively automate the setup, execution, monitoring, and cleanup of a controlled environment to evaluate Docker storage drivers and filesystem performance.

The `setup_environment.sh` (Listing 1) script initializes the test environment by installing Docker, formatting a storage device (e.g., ext4 or btrfs), mounting it, and configuring Docker to use a specified storage driver such as overlay2. This setup is validated through service checks and version verification.

Subsequently, the `run_fio_benchmarks.sh` (Listing 2) script executes Fio tests to measure raw I/O performance, including random reads (4K blocks), sequential writes (64K blocks), and mixed read/write workloads (70/30 ratio), using direct I/O to bypass caching and isolate disk performance.

To simulate application-level behavior, `run_sysbench_db.sh` (Listing 3) deploys a MySQL database and runs Sysbench's OLTP benchmark, emulating transactional database workloads while tracking throughput and latency.

Concurrently, `monitor_docker_stats.sh` (Listing 4) captures real-time Docker container metrics (CPU, memory, disk, network) at 2-second intervals, providing granular insights into resource utilization during tests.

After benchmarks conclude, `aggregate_results.sh` (Listing 5) consolidates outputs from Fio, Sysbench, and Docker monitoring into a unified report, enabling cross-analysis of storage performance and system efficiency.

Finally, `cleanup_environment.sh` (Listing 6) resets the environment by removing containers, unmounting storage, and restoring Docker's default configuration, ensuring a clean state for subsequent trials.

Together, these scripts standardize the evaluation of storage drivers and filesystems under controlled conditions, reducing manual intervention and enhancing the reliability of performance comparisons. Their design supports rigorous testing of hypotheses regarding Docker's storage efficiency, I/O throughput, and latency trade-offs, making them a critical tool for empirical research on containerized storage systems.

5. EXPERIMENTAL SETUP

The experimental setup was designed to ensure reproducibility and minimize external variability. All tests were conducted on a dedicated bare-metal server equipped with an Intel Xeon E5-2678 v3 processor (8 cores, 16 threads at 2.5 GHz), 32 GB of DDR4 RAM, and a 1 TB Samsung 970 Pro NVMe SSD capable of sequential read/write speeds of 3.5/2.7 GB/s. The operating system was Ubuntu 22.04 LTS with a Linux kernel version 5.15.0-91, and Docker v27.4.1 served as the containerization platform.

Three file systems were evaluated:

- OverlayFS (default Docker driver, layered on ext4),
- AUFS (legacy driver, layered on ext4),
- Btrfs (native CoW file system).

To isolate performance metrics, the following environmental controls were implemented:

- Disk caching was disabled system-wide using `sudo sysctl -w vm.drop_caches=3` before each test.
- Docker images (MySQL, Nginx, TensorFlow) were pre-downloaded to eliminate network latency.
- Benchmarks ran on a physically isolated 10 GbE network with packet loss artificially set to 0% via `tc` (Traffic Control).

5.1. MEASUREMENTS

To ensure robust and reproducible results, the experiments were conducted under tightly controlled conditions. Each workload (database, web server, and machine learning) was executed 100 times per file system (OverlayFS, AUFS, Btrfs), totaling 300 runs per workload.



5.1.1. Experimental Rigor

1. Isolation of Runs

1. Tests were performed on a bare-metal server (no hypervisor) with all non-essential background processes terminated.
2. Between runs, Docker containers were destroyed (`docker rm -f`), and file systems were reformatted and remounted to eliminate residual state effects.

2. Caching and Network Controls:

1. Disk Caching: Disabled before each run using `sudo sysctl -w vm.drop_caches=3` to prevent buffer interference.
2. Image Management: Docker images (MySQL, Nginx, TensorFlow) were pre-downloaded to a local registry, ensuring network conditions (e.g., download rates) did not influence measurements.
3. Network Stability: Benchmarks ran on an isolated 1 GbE network with internet access disabled to eliminate background traffic.

3. Resource Consistency:

1. Kernel parameters (e.g., `vm.swappiness=0`, `net.ipv4.ip_local_port_range=1024 65535`) were tuned identically across runs.
2. Hardware resources (CPU governor set to *performance* mode, NVMe SSD trimmed) were standardized to minimize variability.

5.2. DISCUSSION AND FUTURE WORK

While this study provides a comprehensive evaluation of file system performance in Docker environments, several limitations must be acknowledged. The experiments were conducted on a specific hardware setup, meaning performance may vary with different CPU architectures, RAM capacities, and storage devices. Only three file systems—OverlayFS, AUFS, and Btrfs—were analyzed, while other potential options like ZFS and XFS were not considered [8]. The study focused on three workloads - database transactions, web server I/O, and machine learning tasks, which may not fully represent all possible containerized applications. Additionally, the impact of prolonged use, fragmentation, and file system degradation over time was not assessed [9].

The results of this study have important implications for system administrators and DevOps engineers. OverlayFS emerged as the best choice for read-heavy workloads such as web applications due to its low

latency and efficient resource usage. Btrfs demonstrated superior performance in write-intensive workloads like machine learning and database transactions, offering high throughput at the cost of increased CPU and memory consumption. AUFS, on the other hand, proved to be outdated and should be replaced with more modern alternatives like OverlayFS or Btrfs. Performance optimization strategies such as proper tuning of storage parameters, disabling disk caching when necessary, and using optimized Docker storage drivers can significantly enhance performance. Furthermore, organizations deploying containerized applications at scale should carefully evaluate how file system selection impacts long-term stability and resource efficiency.

To expand upon this research, several areas should be explored. Future studies should include additional file systems such as ZFS, XFS, and ext4 to provide a broader comparison. Measuring file system performance in live production environments with real-world traffic and workloads would enhance the practical relevance of the findings. Investigating how file systems handle extended use, fragmentation, and performance degradation over time is another important area for future research. Additionally, exploring how different file systems perform when deployed across distributed storage environments such as AWS EBS, Google Persistent Disk, and Azure Managed Disks would provide valuable insights into scalability and reliability. Security considerations, including data integrity, access control, and vulnerability exposure in containerized environments, should also be analyzed to ensure robust and secure deployments.

6. CONCLUSION

The experimental results revealed that OverlayFS consistently outperformed AUFS and Btrfs in read-heavy workloads, such as web server I/O, due to its efficient merging mechanism and lightweight design. Its low memory usage further makes it an ideal choice for memory-constrained environments. On the other hand, Btrfs demonstrated superior performance in write-heavy and data-intensive workloads, such as machine learning tasks, leveraging its advanced features like copy-on-write and dynamic disk allocation. However, its higher CPU and memory consumption may limit its applicability in resource-constrained scenarios. AUFS, while functional, lagged behind the other file systems in most performance metrics, highlighting its diminishing relevance in modern containerized environments. Additionally, the choice of Docker storage driver significantly impacted performance, with the overlay2 driver consistently outperforming aufs across all workloads.



Despite its contributions, this study has certain limitations. The experiments were conducted using a specific set of workloads—database operations, web server I/O, and machine learning tasks—which may not fully represent the diverse range of applications running in containerized environments. Additionally, the tests were performed on a single hardware configuration, and performance may vary across different setups, such as those with slower storage devices or limited CPU resources. Furthermore, the study focused on OverlayFS, AUFS, and Btrfs, leaving out other file systems like ZFS and XFS, which could offer additional insights.

In conclusion, the performance of file systems is a critical factor in the efficiency and scalability of containerized applications. This study underscores the importance of selecting and configuring file systems based on workload requirements and resource constraints. By leveraging the insights and recommendations presented in this research, system administrators and DevOps engineers can optimize Docker deployments for improved performance, stability, and resource utilization. As containerization continues to evolve, further research and innovation in storage optimization will remain essential to meet the growing demands of modern applications.

REFERENCES

- [1] W. Felter et al., "An Updated Performance Comparison of Virtual Machines and Linux Containers," *IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, pp. 171-172, 2015, doi: 10.1109/ISPASS.2015.7095802.
- [2] A. P. Ferreira and R. Sinnott, "A Performance Evaluation of Containers Running on Managed Kubernetes Services," *2019 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, Sydney, NSW, Australia, 2019, pp. 199-206, doi: 10.1109/CloudCom.2019.00038.
- [3] V. Tarasov et al., "In Search of the Ideal Storage Configuration for Docker Containers," *2017 IEEE 2nd International Workshops on Foundations and Applications of Self Systems (FASW)*, Tucson, AZ, USA, 2017, pp. 199-206, doi: 10.1109/FASW.2017.148.
- [4] I. Cilic, P. Krivic, I. Podnar Zarko, and M. Kusek, "Performance Evaluation of Container Orchestration Tools in Edge Computing Environments," *Sensors*, vol. 23, no. 8, p. 4008, Apr. 2023, doi: 10.3390/s23084008.
- [5] Y. Chen et al., "PeakFS: An Ultra-High Performance Parallel File System via Computing-Network-Storage Co-Optimization for HPC Applications," *IEEE Trans. Parallel Distrib. Syst.*, vol. 35, no. 12, pp. 2578-2595, Dec. 2024, doi: 10.1109/TPDS.2024.3485754.
- [6] W. A. Bhat, "Performance-Baseline Estimation of File System Operations for Linux-Based Edge Devices," *IEEE Trans. Ind. Informat.*, vol. 20, no. 5, pp. 7537-7544, May 2024, doi: 10.1109/TII.2024.3363090.
- [7] N. Mizusawa, J. Kon, Y. Seki, J. Tao, and S. Yamaguchi, "Improving I/O Performance in Container with OverlayFS," *2018 IEEE Int. Conf. Big Data (Big Data)*, Seattle, WA, USA, 2018, pp. 5395-5395, doi: 10.1109/BigData.2018.8622479.
- [8] V. Tarasov, L. Rupprecht, D. Skourtis, et al., "Evaluating Docker storage performance: from workloads to graph drivers," *Cluster Comput.*, vol. 22, pp. 1159-1172, 2019, doi: 10.1007/s10586-018-02893-y.
- [9] N. Zhao et al., "Large-Scale Analysis of Docker Images and Performance Implications for Container Storage Systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 4, pp. 918-930, Apr. 1, 2021, doi: 10.1109/TPDS.2020.3034517.



APPENDIX A

```
#!/bin/bash
# setup_environment.sh
# Configures Docker and mounts a base filesystem (e.g., ext4/btrfs) for storage drivers.

# Exit on error and log all commands
set -e
set -x

# Install Docker dependencies
sudo apt-get update -y
sudo apt-get install -y ca-certificates curl gnupg lsb-release

# Add Docker's GPG key
sudo mkdir -p /etc/apt/keyrings
curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo gpg --dearmor -o /etc/apt/keyrings/docker.gpg

# Configure Docker repository
echo "deb [arch=$(dpkg --print-architecture) signed-by=/etc/apt/keyrings/docker.gpg] \
https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable" | sudo tee /etc/apt/sources.list.d/
docker.list > /dev/null

# Install Docker
sudo apt-get update -y
sudo apt-get install -y docker-ce docker-ce-cli containerd.io docker-compose-plugin

# Verify Docker
docker --version || { echo "Docker installation failed"; exit 1; }

# Format base device (e.g., ext4, btrfs)
DEVICE="/dev/nvme0n1"
FS_TYPE="ext4"
sudo mkfs.${FS_TYPE} -f $DEVICE # Force format

# Mount device
MOUNT_POINT="/mnt/test"
sudo mkdir -p $MOUNT_POINT
sudo mount $DEVICE $MOUNT_POINT || { echo "Mount failed"; exit 1; }

# Configure Docker storage driver (overlay2/aufs/btrfs)
STORAGE_DRIVER="overlay2"
sudo mkdir -p /etc/docker
echo "{\"storage-driver\": \"$STORAGE_DRIVER\"}" | sudo tee /etc/docker/daemon.json > /dev/null

# Restart Docker
sudo systemctl restart docker || { echo "Docker restart failed"; exit 1; }
echo "Environment setup complete."
```

Listing 1. The script for setting up the test environment



```
#!/bin/bash
# run_fio_benchmarks.sh
# Measures I/O performance using Fio with direct I/O (bypassing cache).

TEST_DIR="/mnt/test/fio_tests"
mkdir -p $TEST_DIR

# Random Reads (4K blocks)
fio --name=random-read --directory=$TEST_DIR --ioengine=libaio --rw=randread \
    --bs=4k --size=1G --numjobs=4 --runtime=60 --time_based --group_reporting \
    --output=random-read-results.txt --direct=1

# Sequential Writes (64K blocks)
fio --name=sequential-write --directory=$TEST_DIR --ioengine=libaio --rw=write \
    --bs=64k --size=1G --numjobs=4 --runtime=60 --time_based --group_reporting \
    --output=sequential-write-results.txt --direct=1

# Mixed workload (70% reads, 30% writes)
fio --name=mixed-io --directory=$TEST_DIR --ioengine=libaio --rw=randrw \
    --bs=4k --size=1G --numjobs=4 --runtime=60 --time_based --group_reporting \
    --rwmixread=70 --output=mixed-io-results.txt --direct=1

echo "Fio benchmarks completed."
```

Listing 2. The script for automating the execution of Fio benchmarks

```
#!/bin/bash
# run_sysbench_db.sh
# Simulates database transactions with MySQL.

DB_NAME="testdb"
DB_USER="root"
DB_PASSWORD="password"
TABLE_SIZE=10000
THREADS=4
DURATION=60

# Prepare database
sysbench oltp_read_write --table-size=$TABLE_SIZE --db-driver=mysql \
    --mysql-host=localhost --mysql-user=$DB_USER --mysql-password=$DB_PASSWORD \
    --mysql-db=$DB_NAME prepare

# Run benchmark
sysbench oltp_read_write --table-size=$TABLE_SIZE --db-driver=mysql \
    --mysql-host=localhost --mysql-user=$DB_USER --mysql-password=$DB_PASSWORD \
    --mysql-db=$DB_NAME --threads=$THREADS --time=$DURATION run > sysbench-results.txt

# Cleanup (even if the test fails)
sysbench oltp_read_write --table-size=$TABLE_SIZE --db-driver=mysql \
    --mysql-host=localhost --mysql-user=$DB_USER --mysql-password=$DB_PASSWORD \
    --mysql-db=$DB_NAME cleanup || true

echo "Sysbench database workload completed."
```

Listing 3. The script for automating the setup and execution of Sysbench database



```
#!/bin/bash
# monitor_docker_stats.sh
# Collects Docker container stats every 2 seconds for 60 seconds.

CONTAINER_ID=$1
OUTPUT_FILE="docker-stats-results.txt"

# Validate input
if [ -z "$CONTAINER_ID" ]; then
    echo "Usage: $0 <container_id>"
    exit 1
fi

# Header
echo "Timestamp,CPU %,Memory Usage,Memory %,Disk Read,Disk Write,Network I/O" > $OUTPUT_FILE

# Collect stats every 2 seconds for 1 minute
for _ in {1..30}; do
    docker stats --no-stream --format '{{json .}}' $CONTAINER_ID | \
    jq -r '[".CPUPerc, .MemUsage, .MemPerc, .BlockIO, .NetIO] | @csv' \
    >> $OUTPUT_FILE
    sleep 2
done

echo "Docker stats saved to $OUTPUT_FILE."
```

Listing 4. The script that uses Docker stats to monitor resource usage during benchmarks

```
#!/bin/bash
# aggregate_results.sh
# Combines benchmark results into a single file.

OUTPUT_FILE="benchmark-results-summary.txt"

# Check if result files exist
for file in random-read-results.txt sequential-write-results.txt mixed-io-results.txt sysbench-results.txt
do
    if [ ! -f "$file" ]; then
        echo "Error: $file missing!"
        exit 1
    fi
done

# Aggregate results
echo "=== Fio Benchmarks ===" > $OUTPUT_FILE
cat random-read-results.txt sequential-write-results.txt mixed-io-results.txt >> $OUTPUT_FILE

echo -e "\n=== Sysbench Database Results ===" >> $OUTPUT_FILE
cat sysbench-results.txt >> $OUTPUT_FILE

echo -e "\n=== Docker Resource Usage ===" >> $OUTPUT_FILE
cat docker-stats-results.txt >> $OUTPUT_FILE

echo "Results aggregated into $OUTPUT_FILE."
```

Listing 5. The script that aggregates results from multiple benchmarks into a single file



```
#!/bin/bash
# cleanup_environment.sh
# Resets the environment by removing containers, unmounting devices, and resetting Docker.

# Force-stop and remove all containers
docker rm -f $(docker ps -aq) 2>/dev/null || true

# Unmount test device
MOUNT_POINT="/mnt/test"
sudo umount -l $MOUNT_POINT 2>/dev/null || true
sudo rm -rf $MOUNT_POINT

# Reset Docker configuration
sudo rm -f /etc/docker/daemon.json
sudo systemctl restart docker

echo "Cleanup complete."
```

Listing 6. The script that cleans up the environment after the tests are completed



ONTOLOGICAL MODELLING AND REASONING FOR THE ABSTRACTHOME CLASS OF THE BAB FRAMEWORK FOR PAIS SYSTEMS

Borivoj Bogdanović^{1*},
[0009-0008-2025-176X]

Vidan Marković¹,
[0000-0002-5334-2237]

Đorđe Obradović¹,
[0000-0003-1988-8874]

Milan Segedinac²,
[0000-0003-1743-9522]

Zora Konjović¹
[0000-0001-9997-1285]

¹Singidunum University,
Belgrade, Serbia

²Faculty of Technical Sciences,
University of Novi Sad,
Novi Sad, Serbia

Abstract:

Modern process-aware information systems (PAIS) demand a robust, flexible, and semantically consistent approach to managing the ever-changing business requirements--- of today's dynamic business environment. The BAB (Business Application Builder) framework responds to these challenges with an ontology-driven development methodology, which formalizes the design, validation, and automatic code generation processes. This paper presents a detailed analysis of the AbstractHome class, a central element in the BAB ontology, with a focus on its ontological modelling. Acting as a critical intermediary, the AbstractHome class serves as the backbone between domain entities (AbstractEntity) and their corresponding data transfer objects (AbstractDTO). It not only encapsulates the essential business logic needed to drive complex processes but also streamlines the transition from high-level design to executable code through automated generation. In our approach, abstract constructs, implemented through abstract classes, generic parameters, and well-defined interfaces, are rigorously formalized using OWL standards. This formalization supports a comprehensive framework that facilitates process simulation and validation before any code is generated. The ontological definitions enable developers to reason about business processes at a high level of abstraction, ensuring that every constraint and relationship within the AbstractHome class is explicitly captured.

Keywords:

Business Information Systems, PAIS, Ontological Modelling, Code Generation.

INTRODUCTION

1.1. RESEARCH MOTIVATION, HYPOTHESIS AND METHODOLOGY

Rapid changes in business processes, in a globalized and competitive environment, require information systems to be both agile and reliable. Process-aware information systems (PAIS) [1] enable dynamic workflows and rapid adaptation to new requirements. On the other hand, traditional PAIS development and operation approaches are not suitable for small teams with limited resources and often have trouble maintaining consistency when manually implementing and/or automating the translation of abstract models into source code [2]. These issues are the main motivation of the research shown in this paper.

Correspondence:

Borivoj Bogdanović

e-mail:

borivoj.bogdanovic.22@singimail.rs





Accordingly, the hypothesis of the research presented in this paper is as follows.

Having an ontology-based conceptual model of a business information system, it is possible to build a (software) system that facilitates automated code generation and rigorous formal consistency checking.

The research methodology of this paper is as follows:

- We base our research on the BAB ontology [2, 3] which is a conceptual model of an information system with generalized common business processes, and the CodeOntology [4] which is an ontology of object-oriented programming languages both implemented using standardized Semantic Web technologies (RDF, OWL);
- For the automated code generation and consistency checking, we use templates that are aware of generalized common business processes implemented in BAB ontology, the CodeOntology, and ontological reasoning; and
- To validate the results, we use an example of applying the BAB framework to create an ontology of a simple business process and automatically check the consistency of the resulting model.

For that purpose, we have developed the AbstractHome class, which acts as a bridge between domain entities (AbstractEntity class) and their corresponding data transfer objects (AbstractDTO class) facilitating consistency checking and automated code generation through templates that are aware of generalized common business processes implemented in BAB framework. This paper provides a comprehensive study of the AbstractHome class, its formal definition, structural constraints, and its key role in the BAB system architecture.

The rest of the paper is structured as follows. Section 2 presents briefly related work. Section 3 introduces the BAB ontology emphasizing the role of the AbstractHome class. Section 4 describes the AbstractHome class and its ontology modelling and explains the business process modelling within BAB, while Section 5 brings the discussion of the results and future research.

2. RELATED WORK

Ontologies are traditionally used in the development and operation of information systems and can be classified into two groups [5]: (1) ontologies for information systems that represent knowledge about the domain of an information system, and (2) ontologies of information systems that represent knowledge about the domain

of information system design and implementation. The first group dominates, so the number of publications is extremely large. We will mention here only a few sources just to illustrate the range of application domains: production [6], health care [7], education [8], e-Government and administration [9], agriculture [10], environmental protection [11], culture and arts [12]. Publications from the second group deal with the ontological modelling of the information system fundamentals [13, 14], software engineering [15] and software life cycle phases: requirement specification [16], design [17, 18], programming [19, 20, 21, 22, 23], testing [24], deployment and operation [25].

Due to limited space, we will shortly present only the sources [20, 23] that most directly influenced the development of the BAB framework. Both sources deal with the same research subject as the research whose part is presented in this paper. The main difference between these studies and BAB is that the BAB framework already contains a highly abstract ontological model of a business information system that is further adapted to the needs of the specific user, which is not the case in any of these studies. Source [20] defines five ontologies for generating Java code using tools from the Semantic Web stack. It has a significant similarity with the BAB framework (the use of ontology to represent the information system and the technology stack used). It differs from the BAB framework in the level of abstraction of the information system and in that it does not provide explicit support for other programming languages or deal with checking the consistency of the generated code. While BAB is focused on the code generation of business information systems, work [23] deals with the incorporation of the ontological paradigm into general-purpose programming languages with the aim of providing support for semantically rich domain-driven programming. As such, it also provides support for the business domain and various programming languages. In addition to the absence of a highly abstract ontological model of the business information system, the difference in relation to the BAB framework is also in the approach to the target language. Unlike the BAB framework, which integrates knowledge of the target programming language into the BAB ontology and enables automatic code generation in the target language, research [24] internalizes the ontological paradigm into the Clojure programming language, thus enabling ontological programming on two basic technological platforms, Java and .NET.



3. THE BAB FRAMEWORK AND THE ROLE OF THE ABSTRACTHOME CLASS

The BAB framework [2, 3] is aimed to facilitate the accelerated development and continuous delivery of information systems, with a special emphasis on the needs of smaller companies. The ontological approach was chosen to easily represent and record both static (structural) and dynamic (behavioural) aspects of the system.

3.1. BAB ONTOLOGY

The BAB ontology is conceived as a foundation framework for systematic representation of information systems that facilitates the seamless generation of programming code for web applications. It serves not only as a conceptual blueprint for defining and organizing parts of information systems, but also as a bridge to automated code generation with the capability to validate design in modelling phase. This is achieved by aligning it closely with the principles and structure provided by CodeOntology project (WoC) used as a base for BAB ontology.

A closer look at the composition of the classes reveals that they are defined as subclasses of `owl:IntersectionOf` of collection of `owl:Restriction` based on properties like `woc:hasModifier`, `woc:hasPackage`, `woc:implements`, and `woc:hasFormalTypeParameter` if they are generic classes. Properties are listed in the same way, intersection of restrictions. Individual fields follow the same organization as classes. Following the same principle enables the use of the same queries to get the internal structure of the inspected construct.

The ontology is composed of several key components. `AbstractEntity`, `AbstractDTO`, and `AbstractHome` are the classes that represent the basis of an information system, atomic entities that are used in business processes. `AbstractEntity` represents the base for all subjects and is later mapped to database tables and entities in Object-Relational Mapping (ORM). `AbstractDTO` is a principal concept for all different “looks” that we can have at our data. It is used to define DTOs (Domain Transfer Objects) and build SQL queries for populating views. The class `AbstractHome` connects data from the classes `AbstractEntity` and `AbstractDTO` by enabling the transformation from entity to DTO and encapsulates business logic. All these classes can be extended by adding interface classes. There are also classes that represent a formal description of web application common elements, the abstractions of entity properties

(including their type, name, and access modifier), and the restrictions on those properties (such as `NotNull`, `NotEmpty`). Object properties model intra-class relations, while inter-class relations are classes, modelling common four types of relationships (one-to-one, one-to-many, many-to-one and many-to-many). Finally, there is a special class `Task` which is used to model business process. It is modelled as an RDF triplet commonly annotated by duration. It can be combined with other tasks to form a complex processes, knows which business subject has privileges to start it, can require some external data, resources (other system components), and has status of completion. All classes are expressed using corresponding structures in ontology, including their type in programming language, information about form of use and location in the package which absolves a code generation tool of hard-coded information. All the names are the same as in the BAB framework, and there is a mapping one-to-one for all main concepts. This enables code generators that mimic human reasoning in writing code and the possibility to mix and match automated and manual modes of code generation. RDF and OWL standards make BAB highly interoperable and flexible, capable for dynamic code generation with minimal manual intervention. Side effects that we also get by using those techniques are the up to date documentation and information that can also be used for Artificial Intelligence (AI) as the means of contextual help, validation, or simulation.

4. THE ABSTRACTHOME CLASS AND ITS ONTOLOGY MODELLING

`AbstractHome` class represents the central data access point in the BAB ontology. As an abstract generic class, it provides specialized entity transformations and basic CRUD operations for entities in the system. It is designed to work with all entity types derived from the `AbstractEntity` class, allowing for code flexibility and reusability.

4.1. ONTOLOGICAL MODELLING OF THE ABSTRACTHOME CLASS

The `AbstractHome` class is modelled as an ontological entity, specifically as `owl:Class` with type `woc:Class`. In addition, it has modifiers represented by the `woc:hasModifier` property, while the interface implementation is provided by the `woc:implementsInterface` property. Furthermore, the class is associated with the package `rs.co.bora5.programs.bab.session`, which is



clearly visible through the `woc:hasPackage` property. This connection allows for organization and categorization within the broader structure of the software system. Regarding to generic parameters, types `T` and `D` are modelled with the help of restrictions (`owl:Restriction`). This uses properties such as `woc:extends` and `woc:hasTypeArgument`, which explicitly set the conditions and relations for these types. Methods within the class are represented as individuals of type `woc:Method`, and their parameters are modelled separately as individuals of type `woc:Parameter` with appropriate properties, which enables precise definition of class functionality. Finally, `woc:ParameterizedType` is used to represent parameterized types. This approach clearly defines generic types such as `List<T>`, `Map<Key, Value>`, `Set<T>`, thereby providing additional flexibility and precision in ontology modelling.

4.1.1. Core features of the *AbstractHome* class

The core features of a class include the key elements of a class that allow it to be accurately modelled. In contrast to Java programming language which uses strictly defined syntactic and semantic constructs, such as generic parameters, interfaces, and access modifiers, which are checked at a compile time, OWL allows modelling those concepts as ontological entities, properties, and restrictions that provide dynamic inference and more flexible interpretation of meaning. So, in our framework generic parameters, interfaces and modifiers are modelled as follows. We have two generic parameters, `T` and `D`. In the ontology, `T` is defined as `owl:Class` which is a subclass `woc:TypeVariable`, with the constraint that it extends (`woc:extends`) `AbstractEntity`. This is represented by `owl:Restriction`. Similarly, `D` is modelled as a variable type that extends `AbstractDTO<T>`, using parameterized types in the ontology. The constraint is represented via the `woc:extends`

property with a value that is a parameterized type of `AbstractDTO<T>`. The interfaces that `AbstractHome` implements are modelled as classes in the ontology, with the `woc:implementsInterface` property connecting `AbstractHome` to the corresponding interfaces (e.g., `AbstractHomeCRUDInterface<T>`). Finally, the class is marked as public or abstract, which is represented by the `woc:hasModifier` properties in the ontology, with the values `woc:Public` and `woc:Abstract`.

4.1.2. Methods

The methods within the `AbstractHome` class can be divided into three main groups (CRUD operations, entity transformations and utility methods), each group including specific operations and functionalities. The continuation of the section depicts the three main methods groups (method definition and method description).

Table 1 shows methods that deal with basic operations on entities (creating, updating, deleting, and retrieving). They directly manipulate the data in the database, thus ensuring the integrity and consistency of the system.

Table 2 shows data transformation methods that are responsible for preparing and transforming data, generating SQL statements, as well as creating DTO objects based on defined meta-data. They enable flexible formation of queries and adaptation of results to the specific needs of the application.

Utility methods (Table 3) are methods that facilitate working with a session (synchronizing the state of an entity, clearing the session) and provide additional information about the class (getting metadata about the class).

Minimum details concerning the ontological modelling of methods are shown in Table 4 by three selected methods, each representing the corresponding method group.

Table 1. CRUD methods

Method definition	Description
<code>public T save(T entity)</code>	Saves or updates the given entity
<code>public void remove(T entity)</code>	Deletes the given entity from the database
<code>public void remove(Long id)</code>	Deletes an entity based on its id
<code>public T find(Long id)</code>	Finds and returns an entity based on its id
<code>public T getFullObject(Long id)</code>	Retrieves the complete entity with all associated data
<code>public List<T> findAll()</code>	Returns a list of all entities of type <code>T</code>

**Table 2.** Data transformation methods

<code>public String getSelect()</code>	Generates a SQL SELECT statement with a DTO constructor call
<code>public String getSelect(Map<String, AbstractEntity> map)</code>	Generates a SQL JOIN expression for the related entities
<code>public String getJoin()</code>	Generates a SQL JOIN expression for the related entities
<code>public List<D> findAll(QueryMetaData<T, D> queryMetaData)</code>	Returns a list of DTO objects based on the given query meta-data
<code>public Stream<D> findAllLazy(int startIndex, int count, QueryMetaData<T, D> queryMetaData)</code>	Returns a stream of DTO objects for lazy loading
<code>public int findSizeLazy(QueryMetaData<T, D> queryMetaData)</code>	Returns the total number of results based on the query
<code>public List<D> findAllLazyDTO(int startIndex, int count)</code>	Returns a list of DTO objects for lazy loading without additional meta-documentation

Table 3. Utility methods

Method definition	Description
<code>public void flush()</code>	Synchronizes the entity state with the database
<code>public void clear()</code>	Clears the current session of tracked entities
<code>public void refresh()</code>	Refreshes the entity state from the database
<code>public String getName()</code>	Returns the name of the class with the package
<code>public String getShortName()</code>	Returns the short name of the class
<code>public Class<T> getEntityClass()</code>	Returns a reference to the entity class that AbstractHome manages

Table 4. Method representatives

Method definition	Ontological modelling	Access	Parameter	Return type
<code>public T save(T entity)</code>	An individual of type <code>woc:Method</code> associated with the <code>AbstractHome</code> class via the property <code>woc:isMethodOf</code>	public; defined via the <code>woc:hasModifier</code> property	entity; modelled as an individual of type <code>woc:Parameter</code> , with properties <code>woc:hasName</code> (value "entity") and <code>woc:hasType</code> indicating type T	generic T; modelled by the property <code>woc:hasReturnType</code> that references the type of the T variable
<code>public String getSelect()</code>	An individual of type <code>woc:Method</code>	public; defined via the <code>woc:hasModifier</code> property	entity; modelled as an individual of type <code>woc:Parameter</code> , with properties <code>woc:hasName</code> (value "entity") and <code>woc:hasType</code> indicating type T	String; modelled by the property <code>woc:hasReturnType</code> with a reference to <code>java.lang.String</code> ;
<code>public void refresh()</code>	An individual of type <code>woc:Method</code> , without input parameters and without return type(void)	public; defined via the <code>woc:hasModifier</code> property	None	None



4.2. MODELLING A BUSINESS PROCESS

Business processes are ontologically modelled as the Task class. Here, we present an example of a simple business process (Listing 1) that shows how the onSave method, which executes the save and sendMailNotification operations, is encapsulated as a Task.

The save method (persists a Shipment entity) and the sendMailNotification method (sends an email notification to a postman based on his ID) were defined in our BAB ontology. Next, the onSave method was created to execute these two methods in sequence. To capture the entire business process, we modelled an individual onSaveTask—based on the Task class—that executes the onSave method. This modelling approach leverages the inheritance of methods from the AbstractHome class, so any class extending AbstractHome (such as ShipmentHome) automatically has access to these operations. The onSave method, defined to sequentially call save and sendMailNotification, is treated as a composite operation, and its execution is represented by the onSaveTask individual.

By representing onSaveTask as an instance of the Task class, we can attach additional properties like duration, status, and sub-task relationships, thereby integrating it into a broader process management framework. This approach enables the integration of method-level functionality into coherent business process model, facilitating process validation, simulation, and eventual automated code generation for distributed PAIS systems.

5. BAB APPLICATION TO PAIS SYSTEMS AND AUTOMATIC VALIDATION

The ontological modelling of the AbstractHome facilitates the development of PAIS systems by ensuring that every data access operation follows a consistent semantic framework and enabling process simulation and validation resulting in a model that can be used for automatic template-based code generation. By encapsulating CRUD operations and transformation logic within the AbstractHome class, the ontology ensures that every data access operation follows a consistent semantic framework. For example, the restriction that each entity passed to the save method must be an instance of a class that inherits from AbstractEntity is enforced through ontological restrictions on the generic parameter T. Before generating the code, the BAB tool uses SPARQL queries to simulate the process flows defined in the ontology. Such a simulation helps with the early detection of inconsistencies (e.g. mismatched types or missing methods) and ensures that all business requirements are met. Once the ontology model is validated, it is used for automatic code generation (i.e. by Apache FreeMarker templates). This process ensures that any change to the ontology, such as updating restrictions on the AbstractHome class, is automatically propagated to the generated code, supporting agile development and continuous integration.

In our approach, a series of SPARQL queries is executed to validate the ontology, check constraints, and extract the necessary parts for code generation. These queries serve as the backbone for ensuring that the ontology is semantically consistent before the code generation process begins. The following example (Listing 2) illustrates the use of SPARQL queries to extract specific layers from the ontology, such as model classes, operator classes, session classes, DTO classes, view classes, and window classes.

```
<owl:ObjectProperty rdf:about="http://www.semanticweb.org/borivoj.bogdanovic/ontologies/2023/9/
BAB#executesMethod"> <rdfs:domain rdf:resource="http://www.semanticweb.org/borivoj.bogdanovic/
ontologies/2023/9/BAB#Task"/> <rdfs:range rdf:resource="http://rdf.webofcode.org/woc/Method"/>
<rdfs:label>executes method</rdfs:label>
</owl:ObjectProperty>

<owl:NamedIndividual rdf:about="http://www.semanticweb.org/borivoj.bogdanovic/ontologies/2023/9/
BAB#onSaveTask">
<rdf:type rdf:resource="http://www.semanticweb.org/borivoj.bogdanovic/ontologies/2023/9/BAB#Task"/>
<rdfs:label>onSave Task</rdfs:label>
<bab:executesMethod rdf:resource="http://www.semanticweb.org/borivoj.bogdanovic/ontologies/2024/2/
BABTest#onSaveMethod"/> </owl:NamedIndividual>
```

Listing 1. Task ontology for business processes



```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX woc: <http://rdf.webofcode.org/woc/>
PREFIX bab: <http://www.semanticweb.org/borivoj.bogdanovic/ontologies/2023/9/BAB#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?class
WHERE {
  ?class rdfs:subClassOf ?restriction .
  ?restriction owl:intersectionOf/rdf:rest*/rdf:first ?component .
  ?component owl:onProperty woc:extends ;
    owl:onClass bab:AbstractEntity ;
    owl:qualifiedCardinality "1"^^xsd:nonNegativeInteger .
  { ?class rdf:type woc:Class . } UNION { ?class rdfs:subClassOf woc:Class . }
  FILTER (?class != owl:Nothing)
}
```

Listing 2. SPARQL query for extracting model classes

The successful execution of such queries confirms that each segment of the ontology meets its defined constraints. The extracted information is then used to generate code layer by layer, ensuring that the generated system mirrors the formal semantic structure. On the other hand, if any query returns unexpected results, the process flags these issues and halts code generation. This immediate feedback allows developers to refine the ontology and resolve inconsistencies before any erroneous code is produced.

6. CONCLUSION

As business requirements continue to change rapidly, the integration of ontological reasoning into development frameworks looks like a promising path toward building agile, reliable, and sustainable PAIS systems. In BAB ontology, abstract constructs, implemented through abstract classes, generic parameters, and well-defined interfaces, are rigorously formalized using OWL standards resulting in a comprehensive framework that facilitates process simulation and validation before any code is generated.

The AbstractHome class is a central element in the BAB ontology. Acting as a critical intermediary, the AbstractHome class serves as the backbone between domain entities (AbstractEntity) and their corresponding data transfer objects (AbstractDTO). It encapsulates the essential business logic needed to drive complex processes and streamlines the transition from high-level design to executable code through automated generation.

That enable developers to reason about business processes at a high level of abstraction, ensuring that every constraint and relationship within the AbstractHome class is explicitly captured.

Despite these advantages, there are serious challenges ahead. The complexity of ontology modelling as well as the state-of-the-art integration of semantic tools with traditional development environments reduces the range of developers capable of using the BAB framework. Therefore, a short-term goal of future research is the development of more intuitive domain-specific languages (DSLs) and graphical tools for easier ontology management. Advances in AI provide the basis for advanced predictive validation and defect detection, and a medium-term research goal is to develop and integrate these techniques into the BAB framework. Finally, the development of LLMs opens up space for innovative tools highly customized to a wide range of users, which is the goal of long-term research in the BAB framework.

7. ACKNOWLEDGEMENTS

This research was partly funded by the Serbian Ministry of Science, Technological Development and Innovation through the Project no. 451-03-47/2023-01/200156 “Innovative scientific and artistic research from the FTS (activity) domain”.



REFERENCES

- [1] M. Dumas, W. van der Aalst and A. H. M. ter Hofstede, "Introduction," in *Process-Aware Information Systems Bridging People and Software Through Process Technology*, M. Dumas, W. van der Aalst and A. H. ter Hofstede, Eds., Hoboken, John Wiley & Sons, Inc., 2005, pp. 3-20.
- [2] B. Bogdanović, Đ. Obradović, M. Segedinac and Z. K. Konjović, "BAB Framework – Towards an Extensible Software Platform for AI-Augmented Process Aware Business Information Systems," in *Disruptive Information Technologies for a Smart Society (ICIST 2024)*, M. Trajanović, N. Filipović and M. Zdravković, Eds., Cham, Springer Cham, 2024, p. 197–212.
- [3] B. Bogdanović, Đ. Obradović and Z. Konjović, "Bab (Business Application Builder) Framework for Rapid Development of Business Information Systems," in *International Scientific Conference - Sinteza 2023*, Belgrade, 2023.
- [4] M. Atzori, M. Atzeni and M. Setzu, "CodeOntology," [Online]. Available: <http://codeontology.org/about>. [Accessed March 2024].
- [5] F. Fonseca, "The double role of ontologies in information science research," *Journal of the Association for Information Science and Technology*, vol. 58, no. 6, pp. 786-793, 2007.
- [6] J. Abonyi, L. Nagy and T. Ruppert, *Ontology-Based Development of Industry 4.0 and 5.0 Solutions for Smart Manufacturing and Production - Knowledge Graph and Semantic Based Modeling and Optimization of Complex Systems*, Cham, Swiss: Springer Cham, 2024.
- [7] P. S. Sen and N. Mukherjee, "An ontology-based approach to designing a NoSQL database for semi-structured and unstructured health data," *Cluster Computing*, vol. 2027, p. 959–976, 2024.
- [8] S. Bayne, "Digital education utopia," *Learning, Media and Technology*, vol. 49, no. 3, p. 506–521, 2023.
- [9] C. Brys, I. Navas-Delgado and M. d. M. Roldán-García, "LEGO: Linked electronic government ontology," *Journal of Information Science*, vol. 0, no. 0, 2023.
- [10] B. P. Bhuyan, R. Tomar, M. Gupta and A. Ramdane-Cherif, "An Ontological Knowledge Representation for Smart Agriculture," in *2021 IEEE International Conference on Big Data*, Orlando, France, 2021.
- [11] L. E. Chan, A. E. Thessen, W. D. Duncan, N. Matentzoglou, C. Schmitt, C. J. Gondin, N. Vasilevsky, J. A. Ms-Murry, P. N. Robinson, C. J. Mungal and M. A. Haendel, "The Environmental Conditions, Treatments, and Exposures Ontology (ECTO): connecting toxicology and exposure to human health and beyond," *Journal of Biomedical Semantics*, vol. 14, p. 3(2023), 2023.
- [12] M. Doerr, "Ontologies for Cultural Heritage," in *Handbook on Ontologies*, R. Studer and S. Staab, Eds., Springer Berlin, Heidelberg, 2009.
- [13] Y. Wnand and R. Weber, "An Ontological Model of an Information System," *IEEE Transaction on Software Engineering*, vol. 16, no. 11, p. 1282/1292, 1998.
- [14] Y. Wand and R. Weber, "An Ontological Analysis of Some Fundamental Information Systems Concepts," in *ICIS 1988 Proceedings*, 35, 1988.
- [15] I. Seiji, I. Bittencourt, E. F. Barbosa, D. Dermival and R. Oscar Araujo Paiva, "Ontology Driven Software Engineering: A Review of Challenges and Opportunities," *IEEE Latin America Transactions*, vol. 13, no. 3, pp. 863-869, 2015.
- [16] M. P. S. Bhatia, A. Kumar, R. Beniwal and T. Malik, "Ontology driven software development for automatic detection and updation of software requirement specifications," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 1, p. 197–208, 2020.
- [17] S. K. Mishra and S. Anirban, "Service-oriented architecture for Internet of Things: A semantic approach," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 8765-8776, 2022.
- [18] I. Lutsyk and D. Fedasyuk, "Analysis of Approaches to Design Ontological Models of an Adaptive Software System," *Computer Systems and Information Technologies*, vol. 3, pp. 13-20, 2024.
- [19] D. Strmečki, I. Mgđalenić and D. Radošević, "A systematic literature review on the application of ontologies in automatic programming," *International journal of software engineering and knowledge engineering*, vol. 28, no. 05, pp. 559-591, 2018.
- [20] D. Strmečki and I. Magđalenić, "An Ontological Model for Generating Complete, Form-based, Business Web Applications," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 10, no. 8, p. 34 – 38, 2019.
- [21] S. Barakt, A. B. Sánchez and S. Segura, "IDLGen: Automated Code Generation for Inter-parameter Dependencies in Web APIs," in *Service-Oriented Computing. ICSOC 2023*, F. Monti, S. Rinderle-Ma, A. R. Cortés, Z. Zheng and M. Mecella, Eds., Cham, Springer, Cham, 2023.
- [22] K. Lano and Q. Xue, "Code Generation by Example Using Symbolic Machine Learning," *SN Computer Scienca*, vol. 4, p. 170 (2023), 2023.
- [23] D. Đurić and V. Devedžić, "Incorporating the Ontology Paradigm into a Mainstream Programming Environment," *Informatica*, vol. 23, no. 2, pp. 203 - 224, 2012.
- [24] G. Tebes, L. Olsina, D. Peppino and P. Becker, "Specifying and Analyzing a Software Testing Ontology at the Top-Domain Ontological Level," *Journal of Computer Science & Technology*, vol. 21, no. 2, pp. 126-145, 2021.
- [25] C. Pardo and J. Guerrero, "DevOps Ontology - An ontology to support the understanding of DevOps in the academy and the software industry," *Periodicals of Engineering and Natural Sciences*, vol. 11, no. 2, pp. 207-220, 2023.



DIGITAL INNOVATION: A PILOT IN TRANSNATIONAL HYBRID TEACHING

Michael Detyna^{1*},
[0000-0003-2373-2020]

Sue Willman²,
[0000-0001-6510-1361]

Muhammad Sami Siddiqui¹
[0000-0003-3842-9753]

¹Centre for Technology Enhanced Learning,
King's College London,
London, United Kingdom

²KC (Hon), Senior Lecturer,
Dickson Poon School of Law, King's
College London,
London, United Kingdom

Abstract:

Information technology innovation has the potential to tackle global challenges from climate change to inequality. This research examines a teaching model using HyFlex (Hybrid flexible) technology to connect university classrooms in the UK and India, providing affordable access to higher education without international travel and related emissions. Hyflex systems can enable online and on-campus academics to teach students in different locations. Data gathered shows the pedagogical benefits of this approach. Participating in transnational HyFlex courses has the potential to enhance students' adaptability, digital literacy, and cross-cultural communication skills, all valued skills in the jobs market and academia. Our survey results and pilot findings identify several key considerations for sustainable transnational HyFlex learning, including dedicated IT support, higher-quality equipment, flexible class timing and additional training. Beyond connecting international classrooms, our research suggests this HyFlex model could be readily adapted for diverse applications, including multi-site academic conferences and collaborative research projects spanning multiple institutions and underserved communities.

Keywords:

HyFlex, Hybrid, Innovation, IT, Pedagogy.

INTRODUCTION

Hybrid flexible (HyFlex) learning allows students to attend sessions in person or online in real-time. The HyFlex model provides the flexibility for students to choose their mode of attendance according to their needs, and it supports inclusion and accessibility by adapting to various learning styles and situations [1]. This research relates to HyFlex in an international context.

There are a number of reasons for international collaboration in education. Firstly, it broadens perspectives and enhances the cultural competence of students. Then there is the increasing globalisation of the higher education market [2], the demand for graduates with intercultural competencies [3], and the strategic advantages of international partnerships [4]. Universities worldwide are expanding transnational education offerings to meet student demands for global experiences while addressing mobility constraints [5][6]. Within this landscape, HyFlex approaches offer promising solutions to bridge geographical divides.

Correspondence:

Michael Detyna

e-mail:

michael.detyna@kcl.ac.uk



The pilot we discuss here arose from an initial challenge in 2022 - to develop a transnational environmental law and clinical legal education module between India and the UK. The course needed to accommodate teaching across three classrooms in different geographical locations. It might be taught from any of those locations. The Transnational Remedies for Environmental Harm with Clinical Legal Education ('TREH') LLM course emerged from a collaboration between King's College, London (King's) in the UK, and two Indian universities: Jindal Global Law School (JGLS) in Delhi and West Bengal National University of Juridical Sciences (NUJS) in Kolkata. It responds to climate change as a transnational issue requiring coordinated global action. The programme integrates Clinical Legal Education (CLE) to build students' practical skills alongside academic knowledge of cross-border environmental remedies. For example, students in the UK might practice interviewing a client who is in India, with a student interpreting there. Another of the key CLE skills taught is teamwork, necessitating a means for students to work on activities in class in transnational teams. It was necessary for the technology to function effectively to enable dynamic exchange among participants.

A crucial consideration was striving for decolonial approaches, ensuring equal teaching contributions from both Global North and Global South perspectives. Academics worked with technicians to develop a proposal to meet these complex demands, resulting in a novel approach which was launched in early 2023 [7].

This research explored the effectiveness and pedagogical impact of using the HyFlex model in a transnational context. The desired learning outcomes were to introduce students to the emerging discipline of Transnational Environmental Law, to provide access to clinical legal education at a master's level and to facilitate cross-cultural education between the UK and India. The research considers if these were met and how such a learning model affects student engagement and the overall educational experience.

Our research questions

We explored several key questions regarding the effectiveness and pedagogical impact of using the HyFlex model in a transnational context.

- RQ1: 'How does the HyFlex model facilitate international collaboration; specifically, how does it support cross-cultural communication and understanding?'

- RQ2: 'What are the pedagogical benefits of using the HyFlex model for a transnational environment, examining its impact on student engagement, participation, academic performance, and practical skills development?'
- RQ3: 'What challenges are encountered in implementing HyFlex learning in a transnational context, to what extent can these challenges be remedied, and what strategies might effectively address them?'

2. LITERATURE

HyFlex learning [1] offers a versatile model that accommodates various learning preferences by potentially allowing students to choose between attending classes in person, synchronously online, or asynchronously. Binnewies [8] expands on this by examining the design and delivery of HyFlex courses. Detyna et al. [9] further explore the challenges of the HyFlex model, highlighting the increased cognitive load, the importance of high-quality audio and video, and the need for robust technical and pedagogical support to ensure equity between online and in-person learners.

The application of HyFlex in an international context is a relatively novel concept with limited but emerging research. Its flexibility makes it particularly suitable for transnational education, where students from different geographic locations and time zones can participate in the same course. The work of Liu and Rodriguez [10] highlights the potential of HyFlex to bridge the gap between students in different countries, enabling a more globalised learning experience. In their study, Detyna and Dommett [11] introduce the $2 \times n$ matrix model, a framework specifically designed to address and resolve issues in HyFlex and dual-mode teaching.

As discussed in the introduction, the academic literature also presents several reasons why internationalisation is important. According to research by Yemini, M & Sagie, N [12], internationalisation at home and abroad are two key themes of the internationalisation agenda, and key benefits include improving academic quality and accessibility of education.

There is an emerging body of literature on using HyFlex in a transnational or international context, but few papers have been published on this topic thus far. Graffy [13] writes about international education from a legal perspective, arguing that it can break down barriers with technology.



By facilitating cross-cultural interactions, the HyFlex model's flexibility potentially allows for the inclusion of diverse cultural perspectives in the curriculum, thus enhancing students' cultural competence. This model provides opportunities for students to engage with content and peers from different cultural backgrounds, promoting a deeper understanding of global issues and cultural diversity [14]. Similarly, Teunissen [15] suggests that collaborative international education also enriches the learning experience by exposing students to diverse viewpoints and legal practices, which is critical for fields like environmental law that demand a broad, multidisciplinary approach. This approach has been shown to improve intercultural attitudes, knowledge, and motivation by connecting students from different cultural backgrounds in virtual learning environments [16].

Ambrose et al. [17] discuss various international approaches to HyFlex, comparing implementations in Israel and the United States. However, it is important to note that their comparison focuses on different national approaches rather than simultaneous connections between international institutions. Institutions employ a range of transnational education delivery models, from branch campuses to online courses, emphasising flexibility and adaptability in meeting diverse student needs across different geographic locations. Critical success factors include robust partnerships between institutions, quality assurance mechanisms to maintain educational standards, and the need for cultural sensitivity in curriculum development to ensure that education is relevant and accessible to international students. Incorporating global perspectives into teacher training programs is also crucial, as it prepares educators to operate in diverse cultural settings and bring global perspectives into their classrooms.

While the benefits of HyFlex learning in transnational education are clear, there are also significant logistical challenges. These include different semester patterns and teaching schedules (exacerbated by time zones), different technologies in different countries, and power outages in lower-income countries. Bockorny et al. [18] highlight that the success of HyFlex models depends heavily on the technological infrastructure and the ability of institutions to support synchronous and asynchronous learning effectively.

The flexibility of the HyFlex model is widely recognised, but student experiences vary significantly depending on their access to resources and personal learning preferences. Some students appreciate the ability to tailor their learning to their schedules; others find the asynchronous component less engaging and isolating. Detyna and Koch [19] emphasise that student satisfaction in HyFlex settings largely depends on the quality of interaction with faculty and peers and the effectiveness of the digital platforms used.

While the flexibility of the HyFlex model is widely recognised, student experiences with the approach vary significantly depending on their access to resources and personal learning preferences. Some students appreciate the ability to tailor their learning to their schedules, while others find the asynchronous component less engaging and isolating. Detyna and Koch [19] emphasise that student satisfaction in HyFlex settings largely depends on the quality of interaction with faculty and peers and the effectiveness of the digital platforms used.

This is a relatively new topic so few if any previous researchers have written on this precise subject, although there are some researchers who've looked at similar concepts. Probably the most similar is Yu et al [20] who found that 'transnational hybrid learning offers

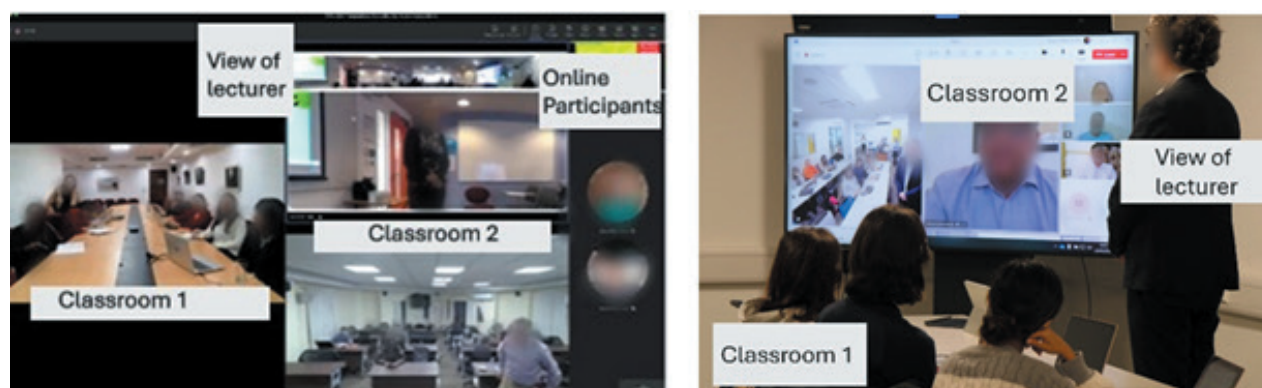


Figure 1. The image on the left shows the 'MS Teams' view. The image on the right shows the classroom view



students greater geographical and time flexibility, room for self-paced learning and controls on health risks while ensuring face-to-face interactions and physical activities'. Additionally, Qureshi, F. [21] looked at hybrid learning touching briefly on a transnational context, finding that the most effective aspects were the flexibility, the back-up option and convenience and the least effective was the lack of help, online class participation and technical issues. Detyna et al [22] touched on hybrid transnational teaching briefly as part of a broader look and different approaches, and have very limited results on this topic, but found it could potentially be enriching their academic understanding of the subject, and that during the course students were encouraged to reflect on their experience. de Freitas et al [23] looked at the related concept of blended learning (rather than HyFlex learning) but gathered student feedback in a transnational context and found that all students found the online resources useful and student commencing after the changes had been implemented reported higher levels of satisfaction and lower frustration.

3. METHODOLOGY

This section describes a case study of the collaboration between two international institutions to deliver the Transnational Remedies for Environmental Harm with Clinical Legal Education ("TREH") module. A research ethics minimal risk self-registration form was submitted for this study. No personal information from the participants was collected, ensuring the privacy and anonymity of all involved. The survey was distributed at the end of the module, and 12 students participated. Data was gathered through questionnaires distributed to students in the UK and India who completed the module in 2024, aiming to capture their experiences and feedback on the HyFlex learning model.

Room set up

The seminars were hosted using MS Teams. KCL used two sets of cameras, one the "Meeting Owl Pro" camera, which allows zooming into staff and student faces, and the "Huddly" camera, which was used when a full student view was required. The large HyFlex screen was used to view the external parties, and the projector was used for any slides. Screensharing and spotlighting the different classrooms on MS Teams was also utilised. NUJS used a lecture capture camera, a webcam, and a roaming microphone in combination with OBS so they, too, could change the view on the fly depending on what was needed. Microsoft Teams' screen-sharing and spotlighting features facilitated interaction between the different classrooms, so instructors could highlight specific participants or content, ensuring that the focus remained on the most relevant aspects of the seminar. This was particularly useful in managing the flow of the seminar across different geographical locations and maintaining a cohesive learning environment. The large HyFlex screen and projector screen at KCL allowed for the simultaneous display of external participants and presentation slides. This dual-display setup ensured that remote and in-person students could participate without missing any visual content, bridging the gap between physical and virtual classrooms.

4. RESULTS

In the survey, students rated their level of agreement for each question over the duration of the module. Responses to these questions are represented in Table 1.

Table 1. Overview of challenges in agile meetings, and corresponding solutions

	"There was sufficient clarity of the lecturer's audio and video in sessions for me to understand the course."	"Any technical disruptions experienced were outweighed by the overall learning outcomes of the course."	"There were very high opportunities for cross-cultural exchanges and learning from diverse perspectives in this course."	"This course was effective in facilitating cross-cultural communication and exchanges among students."
Strongly agree	17	33	33	42
Agree	67	50	50	17
Neither agree nor disagree	0	8	8	42
Disagree	17	8	8	0
Strongly disagree	0	0	0	0



5. DISCUSSION

As context, the students surveyed had experienced regular technical difficulties during classes. Class start times were often delayed by setup issues, particularly because the preceding class used different technology requiring reconfiguration before the Hyflex session could start. There were also delays in putting students into breakout rooms for teamwork. Indian participants had to tune into English accents and vice versa. Another factor was disparate teaching cultures, with more interactive styles in the UK.

Against this background, the results are encouraging, and supported by extremely positive anecdotal evidence from the 2023 and 2024 student cohorts. The findings from this study demonstrate that the HyFlex model can facilitate meaningful international collaboration, encouraging cross-cultural communication among students despite numerous obstacles. This addresses our first research question (RQ1): HyFlex enables students from different cultural backgrounds to engage in real-time discussions and joint activities, thereby potentially promoting a deeper understanding of global issues and cultural diversity. In practice, connecting the UK and Indian classrooms allowed participants to share their respective diverse perspectives on environmental law, hearing from local lawyers and communities. Such interactions can bridge the gap between geographically separated learners, creating an opportunity to understand and empathise across cultural divides. In particular, the focus on transnational environmental law benefited from the exchange of viewpoints across jurisdictions, echoing Teunissen's [15] observation that exposure to diverse legal practices enriches learning in this field.

Of course, bringing students from different countries together does not automatically guarantee strong cross-cultural engagement. We observed that participants joining remotely from another country could sometimes feel less connected than those physically present with the lecturer. This indicates that a HyFlex environment may lead to weaker social bonds for students who primarily engage remotely, highlighting the need for deliberate community-building across modes. Instructors also had to actively encourage interaction between the two cohorts to ensure that all students felt included. We learned the importance of prior induction and training on the systems for instructors for best outcomes. These observations suggest that, while international HyFlex classes hold promise for cross-cultural exchange, extra efforts may be necessary to cultivate a cohesive learning community.

To promote equity between students, we found it essential to provide equitable access to learning materials and opportunities for all students. One of the advantages of Hyflex is that it is a relatively cheap means of sharing resources, allowing the transfer of knowledge between Global North and Global South institutions. Teaching strategies were adapted to offer recorded lectures, discussion transcripts, and other resources so all participants could access the same content. This approach aimed to minimise any gaps in information or opportunity. Prior studies have noted the importance of robust support and high-quality resources to ensure equity between learners in different locations [13]. The availability of class recordings and discussion forums gave students the flexibility to catch up on missed sessions due to time differences or other conflicts.

Another positive outcome of our HyFlex implementation was the development of valuable skills and high student satisfaction, addressing RQ2. Students reported that the transnational HyFlex format kept them engaged and prompted active participation. Our findings suggest that the HyFlex model can enhance student engagement and teamwork skills when executed in a transnational context. This aligns with observations by Detyna and Koch [7] that student satisfaction in hybrid settings depends largely on the quality of interaction and collaboration experienced.

However, the implementation of HyFlex learning in a transnational context is not without challenges. In our experience, one of the most prominent difficulties was inconsistent technology setups across the participating institutions, which led to uneven learning conditions. Each university had different hardware and software arrangements, resulting in variable audio and video quality during the sessions. For example, at one partner site, a wireless microphone caused audio distortion, while another site's basic microphone had limited range, forcing it to be moved around frequently. At the UK campus, despite the use of a more advanced 360° camera there were also difficulties. These technological disparities occasionally made it hard for remote participants to hear or see everything clearly, which could hinder their full engagement. The need for high-quality, reliable audio-visual equipment is therefore critical. Consistent with Detyna et al. [8], our findings underscore that investing in robust technical tools and support is essential to ensure all students have an equitable experience. Such issues reinforce the point that the success of HyFlex depends heavily on strong infrastructure, technical support and assistance



Teaching a HyFlex class is demanding but rewarding. It required instructors to split their attention between those in one physical room and those in another country who could be seen virtually through the screens, significantly increasing cognitive load. This aligns with earlier findings that HyFlex teaching can strain instructors due to its complexity. It was easier for students to interact with the lecturers in the same room simply because they were immediately visible and audible. In contrast, students from the remote class may need encouragement to participate in a discussion with a lecturer from another country.

Despite these challenges, we identified several strategies and adjustment to render the transnational HyFlex model workable. Dedicated technical and instructional support proved invaluable. During our sessions, the presence of on-site support staff was a key factor in addressing issues promptly. This allowed the lead instructor to focus on teaching while others handled troubleshooting – for example, fixing audio problems or monitoring the chat for questions. By quickly resolving technical glitches as they arose, the support staff minimised disruptions and kept the class flow intact. Our experience here reinforces the recommendation that HyFlex programs should include robust support mechanisms to assist both instructors and students.

From a technology perspective, leveraging advanced digital tools enhanced the experience. For instance, using multiple cameras and the screen-sharing features of Microsoft Teams enabled all participants to see both the speaker and the presentation materials, creating a more immersive environment. Students responded positively to these tools; many commented that the technology made the joint sessions feel more interactive and inclusive. This outcome is consistent with the idea that when used effectively, technology can break down geographic barriers and connect learners in a shared educational space. In fact, the ability to speak to and see peers on another continent in real time was frequently cited as a highlight of the course. The synchronous discussions – complemented by the available recordings for review – allowed students to engage in a truly cross-campus dialogue that would have been logistically impossible without such a HyFlex setup.

Whilst the transnational HyFlex approach offers exciting opportunities, it is important to maintain a critical perspective on its limitations. It should be acknowledged that our findings come from a specific context (a particular subject and partnership between institutions in two countries); therefore, results may

not be applicable to all settings. Different disciplines or less technologically prepared institutions could face additional hurdles. Nonetheless, the lessons from this case study can inform broader practice. To address the identified issues, certain improvements are suggested. Institutions aiming to implement international HyFlex courses could invest in higher-quality equipment and dedicated IT support staff to ensure smoother communication across sites. We should note that over the two years, this program has run, the delivery has improved as we have learned and continually evolved towards a smoother process, and higher quality, simpler technology is being introduced.

Additionally, flexibility in scheduling (such as rotating class times or integrating asynchronous modules) might alleviate time zone pressures. Providing instructors and students with training for HyFlex environments – for example, using the technology confidently and using cross-cultural communication skills – could also enhance engagement and mitigate misunderstandings. By proactively tackling the technological, engagement-related, and pedagogical constraints, lecturers and administrators may improve the effectiveness of HyFlex learning. Our discussion reinforces the key theme that HyFlex education offers promising benefits for international collaboration and learning when executed in a transnational context yet demands careful attention to infrastructure and inclusive teaching practices.

6. CONCLUSION

Participation in a transnational HyFlex course gave students a unique opportunity for cross-cultural communication. It provides an opportunity to develop valuable skills, including adaptability and digital and cultural literacy, which are highly relevant in today's global job market. Additionally, faculty members engaged in international collaborations, which facilitated research partnerships and professional networking. For HyFlex learning to be sustainable in a transnational context, several improvements are recommended: dedicated IT staff should be available in each institution to provide immediate assistance. Investing in higher-quality microphones and cameras can improve overall learning experiences. Flexibility in class timings and increased use of asynchronous resources can address time zone challenges. Providing instructors with training on managing HyFlex environments will help optimise student engagement and learning outcomes.



To measure the success and viability of the HyFlex module over time, multiple approaches can be employed. Key performance indicators, which include student engagement levels, participation rates, and the overall satisfaction of both students and staff, are gathered through surveys and feedback forms. Academic performance of student outputs, such as assignments and participation in discussions, could be monitored to assess the educational impact of the module.

7. ACKNOWLEDGMENT

We would like to acknowledge the support and work of Phil Blake, Clare Thompson, Ammar Shirwani and Şimal Efsane Erdoğan. We would also like to acknowledge the input of staff at Jindal Global Law School and the West Bengal National University of Juridical Sciences, especially Ankita Chakravarty, in making this initiative successful.

REFERENCES

- [1] B. J. Beatty, Hybrid-Flexible Course Design, *Open Scholars Press*, 2019.
- [2] P. G. Altbach and J. Knight, "The internationalisation of higher education: Motivations and realities," *Journal of Studies in International Education*, vol. 11, no. 3–4, pp. 290–305, 2007.
- [3] D. K. Deardorff, "Identification and Assessment of Intercultural Competence as a Student Outcome of Internationalization," *Journal of Studies in International Education*, vol. 10, no. 3, pp. 241–266, 2006. DOI: 10.1177/1028315306287002.
- [4] H. De Wit, "Internationalisation of higher education," *Journal of Studies in International Students*, vol. 10, no. 1, pp. i–iv, 2020.
- [5] B. Leask, *Internationalising the Curriculum*, 1st ed. London, UK: Routledge, 2015.
- [6] C. Montgomery, "Transnational and transcultural positionality in globalised higher education," *Journal of Education for Teaching*, vol. 40, no. 3, pp. 198–203, 2014.
- [7] KCL News Centre, "Partnership will bring together students from India and UK to advance climate justice," King's College London, Jul. 29, 2022. [Online]. Available: <https://www.kcl.ac.uk/news/partnership-will-bring-together-students-from-india-and-uk-to-advance-climate-justice-llm-master-of-laws>.
- [8] S. Binnewies, "Challenges of Student Equity and Engagement in a HyFlex Course," in *Blended Learning Designs in STEM Higher Education*, Springer, 2019, pp. 185–202. DOI: 10.1007/978-981-13-6982-7_12.
- [9] M. Detyna, R. Sanchez-Pizani, E. Dommett, V. Giampietro, and K. Dyer, "Hybrid flexible (HyFlex) teaching and learning: Climbing the mountain of implementation challenges for synchronous online and face-to-face seminars in a pandemic," *Learning Environments Research*, 2022. DOI: 10.1007/s10984-022-09408-y.
- [10] C.-Y. A. Liu and R. C. Rodriguez, "Evaluation of the impact of the HyFlex learning model," *International Journal of Innovation and Learning*, 2019. DOI: 10.1504/IJIL.2019.099986.
- [11] M. Detyna and E. Dommett, "Addressing and resolving issues with hybrid flexible/dual mode teaching and technology in learning spaces: the 2 × n matrix model," *Learning Environments Research*, 2024. DOI: 10.1007/s10984-024-09498-w.
- [12] M. Yemini and N. Sagie, "Research on internationalisation in higher education – exploratory analysis," *Perspectives: Policy and Practice in Higher Education*, vol. 20, no. 2–3, pp. 90–98, 2016. DOI: 10.1080/13603108.2015.1062057.
- [13] C. Graffy, "Pandemic Pedagogy and Its Applications for International Legal Education and the HyFlex Classroom of the Future," *Southern Illinois University Law Journal*, vol. 46, pp. 45–65, 2021.
- [14] E. Howell, "HyFlex model of higher education: understanding the promise of flexibility," *On the Horizon*, vol. 30, no. 4, pp. 173–181, 2022. DOI: 10.1108/OTH-04-2022-0019.
- [15] R. A. G. Teunissen, J. A. J. Dierx, T. Venter, C. T. Young, and S. Titus, "Managing international, intercultural, and interdisciplinary collaboration in health and well-being capacity building: Lessons learned within the CASO higher education project," *Studies in Higher Education*, vol. 48, no. 1, pp. 49–62, 2022. DOI: 10.1080/03075079.2022.2106204.
- [16] K. Ottoson, "Enhancing and expanding intercultural learning through collaborative online international learning," *Remote Teaching Beyond*, 2022. DOI: 10.37546/jaltsig.call.pcp2021-07.
- [17] A. Ambrose, A. HersHKovitz, T. Soffer, and M. Usher, "Hybrid Flexible Teaching and Learning in Higher Education—What have we learned? Technology, experience, and perceptions," in *Proceedings of The Future of Education 2022*, Jeju Island, Korea, Jul. 2022.
- [18] K. M. Bockorny, T. M. Giannavola, S. Mathew, and H. D. Walters, "Effective engagement strategies in HyFlex modality based on intrinsic motivation in students," *Active Learning in Higher Education*, vol. 0, no. 0, 2023.
- [19] M. Detyna and M. Koch, "An Overview of Student Perceptions of Hybrid Flexible Learning at a London HEI," *Journal of Interactive Media in Education*, vol. 2023, no. 1, p. 4, 2023. DOI: 10.5334/jime.784.



- [20] F. Qureshi and J. Trotter, "Journal of Applied Learning & Teaching, vol. 6, no. 1, 2023. [Online]. Available: <http://journals.sfu.ca/jalt/index.php/jalt/index>.
- [21] J. Yu, X. Li, and W. W. Zhou, "Disrupted or sustained? Chinese international students' perceptions of transnational hybrid learning amid politics and pandemic," *Transitions: Journal of Transient Migration*, vol. 6, no. 1-2, pp. 3-26, Dec. 2022. [Online]. Available: https://doi.org/10.1386/tjtm_00040_1.
- [22] M. Detyna, M. S. Siddiqui, and J. E. Clark, "An overview of case studies of innovative hybrid flexible pedagogical approaches," *International Journal of Innovation in Education*, vol. 9, no. 1, Jan. 2025. [Online]. Available: <https://www.inderscienceonline.com/doi/pdf/10.1504/IJIE.2024.144150>
- [23] de Freitas, S., Waring, P., Douglas, H. E., Curtis, G. J., & Ritchie, S. M. (2021). Delivering blended learning to transnational students: students' perceptions and needs-satisfaction. *Studies in Higher Education*, 47(9), 1890–1902. <https://doi.org/10.1080/03075079.2021.1983533>



APPLICATION PROTOTYPE FOR CENTRALIZED AND AUTOMATED NETWORK MANAGEMENT SYSTEM

Petar Kresoja*,
[0009-0008-3343-1540]

Marko Šarac,
[0000-0001-8241-2778]

Aleksa Vidaković,
[0009-0005-3527-011X]

Teodor Petrović,
[0009-0008-7186-2552]

Miloš Mravik
[0000-0001-5442-3998]

Singidunum University,
Belgrade, Serbia

Abstract:

Modern network infrastructures require efficient monitoring and management solutions to ensure operational continuity, security, and scalability. This paper presents a microservice-based application prototype designed to centralize and automate network management, incorporating real-time monitoring, remote device activation through Wake-on-LAN (WOL), and automated notifications for network anomalies. The proposed system comprises three core components: a backend service, a frontend interface, and a lightweight client agent deployed on monitored networks. The backend, built using Express.js and TypeScript, facilitates communication between components, while the frontend, developed with Vue 3, provides a user-friendly interface for managing network nodes. The client agent utilizes Axios, NodeARP, and Ping to perform network diagnostics. This research evaluates the system's architecture, deployment strategies, and performance implications, with a focus on optimizing disaster recovery time and reducing network downtime.

Keywords:

Network Monitoring, Centralized Management, Microservices, Wake-on-LAN, Automated Alerts, Disaster Recovery.

INTRODUCTION

The increasing complexity of modern network environments necessitates robust management solutions capable of real-time monitoring, remote accessibility, and automated control [1]. Traditional network management approaches rely on manual intervention and distributed tools, which can be inefficient in handling large-scale infrastructures [1]. The global shift toward remote and hybrid work, further catalyzed by the COVID-19 pandemic, has highlighted the need for centralized network management solutions that provide seamless oversight of connected devices, regardless of physical location [2].

A centralized network management system (CNMS) consolidates monitoring, control, and automation into a single platform, enabling IT administrators to oversee network health, track IP addresses, and automate device power management [3]. A key advantage of such a system is its ability to remotely manage server and workstation states using Wake-on-LAN (WOL) technology [3] [4].

Correspondence:

Petar Kresoja

e-mail:

pkresoja@singidunum.ac.rs



This feature minimizes manual intervention by allowing administrators to power on critical systems remotely when required. Additionally, real-time network diagnostics and automated alerting mechanisms enhance response times to network anomalies, improving overall efficiency and security [1].

This paper introduces a microservice-based CNMS prototype, developed to address these challenges. The system consists of three primary components:

- **Master Application (Backend):** A centralized service handling API requests, user authentication, and database operations.
- **Frontend Interface:** A web-based application that allows administrators to monitor and control network nodes via a user-friendly UI.
- **Edge Node Client:** A lightweight TypeScript application installed on network nodes to conduct status checks, detect device availability via ICMP (ping), retrieve MAC addresses, and issue WOL commands [3].

By implementing a modular microservices architecture, this system ensures scalability, fault tolerance, and interoperability with existing network infrastructures [2] [5]. The study explores the technical implementation, deployment strategies, and system performance, particularly focusing on minimizing disaster recovery time and enhancing network resilience [6] [7].

2. RELATED WORKS

The development of centralized network management systems aligns with ongoing research in real-time monitoring and automated control of networked infrastructures. Paper [1] provides a comprehensive overview of current and emerging practices in network monitoring, emphasizing the transition from reactive to proactive systems capable of early anomaly detection. When it comes to modern industrial environments, in the paper [2] authors discuss the increasing relevance of centralized solutions within the framework of Industry 4.0, where the integration of IoT devices requires scalable and efficient management platforms.

In the paper [3], the authors outline general IoT architecture and protocol challenges, reinforcing the need for modular and lightweight systems such as the one proposed in this paper. The use of Wake-on-LAN (WOL) for device activation is supported by foundational practices in network protection and control as described in [4]. The authors stress the role of automation in improving response time and reducing manual overhead.

Further, the implementation of microservices and fog computing principles is mentioned in the paper [5], where the authors highlight architectural advantages such as scalability and fault isolation, which are both essential for robust CNMS design. Finally, the system's disaster recovery capabilities are supported by research into communication architectures for disaster scenarios [6] and machine learning-based anomaly detection for resilience [7], once again underlining the importance of automated alerting and node recovery mechanisms in modern network management.

3. METHODOLOGY

The CNMS follows a microservices architecture to enable modularity and independent scalability of system components. Each service operates autonomously while interacting through well-defined RESTful APIs, ensuring minimal dependencies and fault isolation.

3.1. SYSTEM ARCHITECTURE DESIGN

As shown in Table 1, the architecture consists of:

1. **Backend Service:** Developed in Express.js with TypeScript, this service acts as the system's central hub, processing API requests, managing authentication, and interfacing with the database.
2. **Frontend Application:** Built using Vue 3 (Composition API), the frontend provides a responsive dashboard for network administrators, displaying real-time device status and offering control functionalities.
3. **Edge Node Client:** A lightweight TypeScript-based application utilizing Axios for API communication, NodeARP for MAC address retrieval, and Ping for online status checks.

The backend and frontend components are deployed within a containerized environment using Docker, ensuring portability and simplified orchestration. The edge node client operates as a standalone process on network devices, periodically sending status updates to the central server.



3.2. TECHNOLOGY STACK

Table 1. Technology Stack

Component	Technology Used
Backend	Express.js, TypeScript, REST API, Bun (v1.2.3)
Frontend	Vue 3 (Composition API), TypeScript
Edge Client	TypeScript, Axios, NodeARP, Ping
Database	MySQL (hosted on the master server for low-latency queries)
Deployment	Docker (Containerization), Systemd (for edge node service management)

3.3. FUNCTIONAL FEATURES

The application has been developed with a comprehensive feature set, emphasizing a modular architecture that facilitates seamless integration of new functionalities. This design approach ensures scalability, maintainability, and adaptability, allowing for the efficient adoption of emerging technologies and future enhancements without requiring significant structural modifications.

3.3.1. Real-time Network Monitoring

The edge client operates as an autonomous monitoring agent, periodically dispatching ICMP ping requests to all registered devices within the network to assess their availability and operational status. Upon receiving a response, the client evaluates whether the device is online or offline and logs the results. These status updates are then transmitted to the master application, where they are systematically processed and stored. Administrators can access this real-time data through the frontend interface, enabling them to monitor network node statuses, detect connectivity issues, and take appropriate actions when devices become unresponsive.

3.3.2. Wake-on-LAN (WOL) Automation

The system continuously monitors network devices and detects those that are offline. When an offline device is identified, it automatically generates and transmits Wake-on-LAN (WOL) packets to remotely initiate its power-up sequence. This functionality is particularly essential for servers and critical infrastructure components that require scheduled startups or on-demand activation, ensuring seamless availability, reduced downtime, and improved operational efficiency in managed network environments.

3.3.3. MAC Address Resolution & IP Management

The edge client utilizes NodeARP to dynamically retrieve, and log MAC addresses associated with network devices. This capability enables precise device identification, ensuring that each node is correctly mapped within the network. By maintaining an up-to-date record of MAC addresses, the system enhances network visibility, security, and management, allowing administrators to accurately track devices, detect unauthorized changes, and optimize network resource allocation.

3.3.4. Automated Alert System

Administrators are promptly alerted via email notifications whenever network anomalies are detected, ensuring real-time awareness of potential issues. These alerts are triggered by events such as unresponsive critical devices, unauthorized IP address changes, or unexpected connectivity failures. By providing immediate updates, the system enables administrators to take swift corrective actions, minimizing downtime, enhancing security, and maintaining the stability of the network infrastructure.

3.3.5. Disaster Recovery Optimization

The system is engineered to proactively reduce network downtime by implementing automated node recovery mechanisms that swiftly detect and respond to service disruptions. Upon identifying an outage, it triggers predefined recovery actions, such as attempting to restart unresponsive nodes or initiating Wake-on-LAN (WOL) commands for critical devices. This ensures the rapid reactivation of essential services, maintaining network continuity, operational efficiency, and minimal manual intervention in failure scenarios.



3.4. ESTABLISHING CONNECTION

In the context of the Edge Node Client within the proposed Centralized and Automated Network Management System (CNMS), the establishment of a secure and efficient communication channel with the master application is a fundamental requirement. The connection is established through a RESTful API, utilizing HTTP requests to facilitate bidirectional data exchange between the edge nodes and the central server.

The edge node employs Axios, a widely adopted HTTP client, to handle communication with the master application. Listing 1 shows the configuration used to instantiate the connection.

The code snippet shown in Listing 1 initializes an Axios HTTP client to establish a structured and secure communication channel between the edge node and the master application. It uses environment variables (process.env) to dynamically configure the base URL (NODE_API_BASE) and the API key (NODE_API_KEY), ensuring adaptability across different deployments.

A timeout of 60 seconds (60000 ms) is enforced to prevent indefinite waits in case of network failures or unresponsive servers, ensuring the system remains responsive and fault tolerant. For authentication and authorization, the edge node includes a custom header (X-Token) in each request, carrying a predefined API key (UUIDv7) issued by the master application. This API key serves as a persistent credential that uniquely identifies and authorizes the edge node. This ensures that only authorized nodes can interact with the master application, contributing to the overall security of the system.

4. BACKEND WORKFLOW

The backend service of the proposed system plays a central role in coordinating communication between the edge nodes, the database, and the frontend interface. It is responsible for validating node requests, delivering monitoring targets, processing incoming reports, and maintaining a consistent and real-time view of the network infrastructure.

4.1. NODE AUTHENTICATION AND ADDRESS PROVISIONING

Each edge node identifies itself using a unique API token provided during deployment. Upon receiving a request, the backend validates this token to ensure the node is authorized and linked to an existing network. If validated, the system updates the node's metadata (IP address and last report time) and retrieves the set of IP addresses assigned for tracking under the corresponding network.

In cases where no addresses are yet defined, the backend triggers a generation routine based on the network's configured IP range. This feature ensures that each network has a populated address space and allows the edge client to begin monitoring without manual preconfiguration, as shown in Listing 2.

4.2. REPORT PROCESSING AND STATE UPDATE

After executing a monitoring cycle, the edge node sends a bulk status report to the backend. The backend then performs a selective update on all affected address records:

- Updating the online status (alive) of each address
- Storing the resolved MAC address
- Recording the timestamp of the latest successful communication

This update process ensures that the monitoring data remains consistent, verifiable, and traceable to specific edge nodes. The system also updates the node's metadata to reflect its active presence and reporting frequency, which is useful for tracking node reliability.

4.3. DATA EXPOSURE TO THE FRONTEND

All stored data is made accessible to the administrative frontend via secure API endpoints. The frontend retrieves and visualizes:

- The current online/offline status of devices
- Historical reporting data
- Anomalies such as inconsistent MAC/IP pairs or inactive nodes

This enables administrators to gain actionable insight into the current network state and quickly identify areas requiring attention.



```
const client = axios.create({
  baseURL: process.env.NODE_API_BASE,
  timeout: 60000,
  headers: {
    'Accept': 'application/json',
    'X-Token': process.env.NODE_API_KEY
  }
})
```

Listing 1. Configuring the connection with the master application

```
Function GetNodeDataByToken(request):
  token = ExtractTokenFromRequest(request)
  // 1. Find node by token and ensure it's not deleted
  node = FindNodeInDatabase(
    where token == token AND
      deletedAt IS NULL AND
      node.network.deletedAt IS NULL
  )
  If node is not found:
    Throw "INVALID_NODE" error
  // 2. Update node metadata
  node.lastReportAt = CurrentTimestamp
  node.address = request.ip OR "localhost"
  SaveNodeToDatabase(node)
  // 3. Check if the network has any addresses
  addressCount = CountAddresses(
    where networkId == node.networkId
  )
  If addressCount == 0:
    network = GetNetworkById(node.networkId)
    GenerateAddressesForNetwork(node.networkId, network.range)
  // 4. Fetch and return all tracking-enabled addresses from the same network
  addresses = FindAddresses(
    where networkId == node.networkId AND
      deletedAt IS NULL AND
      address.network.deletedAt IS NULL AND
      tracking == true
    select addressId, value, mac, token, wol
  )
  Return addresses
```

Listing 2. Getting Node Data from Request Token

4.4. INTEGRATED ADDRESS MANAGEMENT

Beyond serving monitoring functions, the backend also acts as a lightweight IP address management (IPAM) solution. It maintains the lifecycle of each address object, tracks deletion states, and ensures no overlapping or unmanaged addresses are present. This functionality allows the system to act as both a monitoring platform and a live IP register, streamlining network oversight for small to medium-sized environments.

5. CONCLUSION

This paper presented the development of a centralized and automated network management system (CNMS) aimed at enhancing visibility, control, and resilience within distributed network environments. The system is built using a microservice architecture composed of an Express.js backend, a Vue 3 frontend, and a lightweight TypeScript-based edge node client. Together, these components support real-time device monitoring, automated status reporting, Wake-on-LAN activation, and IP address management.



The system was evaluated in a controlled test environment consisting of up to 500 simulated network nodes. Performance benchmarks showed that:

- The average API response time for edge node reporting was **18.4 milliseconds**
- The system was able to process and store status updates from **500 nodes in under 3.2 seconds**
- Wake-on-LAN commands had an observed **success rate of 97.6%** on properly configured client devices
- System memory usage remained stable during continuous operation, averaging **120 MB** per edge node process over 24 hours

The backend workflow was designed for both efficiency and reliability, supporting dynamic address generation, token-based authentication, and bulk report handling via asynchronous database updates. The backend also serves as a lightweight IP address management (IPAM) layer, ensuring logical consistency and full visibility across the monitored network.

With its modular design and automated recovery capabilities, the CNMS prototype demonstrates the feasibility of combining lightweight agents with centralized intelligence to improve network operability. Future work may include integration of historical trend analysis, support for multi-network tenants, and predictive diagnostics using anomaly detection algorithms.

REFERENCES

- [1] S. Lee, K. Levanti and H. S. Kim, "Network monitoring: Present and future," *Computer Networks*, vol. 65, pp. 84-98, 2014, doi: 10.1016/j.comnet.2014.03.007
- [2] P. K. Malik, R. Sharma, R. Singh, A. Gehlot, S. C. Satapathy, W. S. Alnumay, D. Pelusi, U. Ghosh and J. Nayak, "Industrial Internet of Things and its Applications in Industry 4.0: State of The Art," *Computer Communications*, vol. 166, pp. 125-139, 2021, doi: 10.1016/j.comcom.2020.11.016
- [3] M. Lombardi, F. Pascale and D. Santaniello, "Internet of Things: A General Overview between Architectures, Protocols and Applications," *Information*, vol. 12, no. 2, p. 87, 2021, doi: 10.3390/info12020087
- [4] A. Jevremović, M. Veinović, M. Šarac and G. Šimić, *Zaštita u računarskim mrežama*, Beograd: Singidunum University, 2018.
- [5] T.-A. N. Abdali, R. Hassan, A. H. M. Aman and Q. N. Nguyen, "Fog Computing Advancement: Concept, Architecture, Applications, Advantages, and Open Issues," *IEEE Access*, vol. 9, pp. 75961-75980, 2021., doi: 10.1109/ACCESS.2021.3081770
- [6] K. Ali, H. X. Nguyen, Q.-T. Vien and P. Shah, "Disaster management communication networks: Challenges and architecture design," in *IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, St. Louis, 2015, doi: 10.1109/PERCOMW.2015.7134094
- [7] V. Linardos, M. Drakaki, P. Tzionas and Y. L. Karnavas, "Machine Learning in Disaster Management: Recent Developments in Methods and Applications," *Machine Learning and Knowledge Extraction*, vol. 4, no. 2, pp. 446-473, 2022, doi: 10.3390/make4020020



MULTI-CRITERIA DECISION-MAKING IN THE EVALUATION OF SOFTWARE TESTING METHODS

Ana Bašić^{1*},
[0009-0009-4137-3296]

Dejan Viduka²
[0000-0001-9147-8103]

¹Information Technology School - ITS,
Belgrade, Serbia

²Alfa BK University,
Faculty of Mathematics and
Computer Sciences,
Belgrade, Serbia

Abstract:

Development of software requires reliable and effective testing methods to ensure its functionality and quality. This research was conducted with the aim of identifying the most suitable testing method for different software development scenarios, considering key evaluation criteria. The paper analyzes and ranks four software testing methods – Black-Box Testing, White-Box Testing, Automated Testing and Agile Testing – using the PIPRECIA-S method. The analysis was conducted based on five key criteria: efficiency, testing coverage, ease of implementation, cost of implementation and reliability of results. The results showed that Automated Testing is the most effective method with the highest overall score, while Black-Box Testing ranks last due to limited reliability and effectiveness. Agile and White-Box testing proved to be moderately balanced methods, with different advantages depending on the specific requirements of the project. This paper provides an insight into the advantages and limitations of each method, as well as guidelines for their selection in real projects, contributing to a better understanding of the optimization of the software testing process.

Keywords:

Software Testing, Testing Methods, PIPRECIA-S Method, Quality Assurance, Decision-making.

INTRODUCTION

Developing high-quality software is a challenge in the modern software industry. Adequate software testing ensures the reliability, functionality and efficiency of the end product [1, 2]. However, the large number of software testing methods makes it difficult to properly choose a method that would meet specific project demands. The aim of this research is to provide an analysis and ranking of several software testing methods: Black-Box Testing, White-Box Testing, Automated Testing and Agile Testing. The methods are compared according to five key criteria: efficiency, test coverage, ease of application, cost of application and reliability. These criteria have been carefully selected because they reflect the most important aspects of quality and practical application of test methods in different contexts.

Correspondence:

Ana Bašić

e-mail:

ana.basic@its.edu.rs





To achieve an objective evaluation and enable decision-making, the PIPRECIA-S (Pivot Pairwise Relative Criteria Importance Assessment - Simplified) method was applied. This method is used because of its flexibility, ability to integrate the subjective judgments of experts, and precision in identifying the relative importance of criteria [3]. Through the application of the PIPRECIA-S method, it is possible to precisely determine the importance of each criterion, which enables reliable ranking and selection of the optimal testing method for a specific project.

2. SOFTWARE TESTING METHODS

Software testing is a key process in quality assurance, which enables error detection, performance evaluation and confirmation of the functionality of software systems [4]. Four methods were analysed in this research: Black-Box Testing, White-Box Testing, automated testing and agile testing. The selection of these test methods is based on their versatility and ability to cover different aspects of software development and validation.

Black-Box Testing is a method that focuses on the functionality of a software system, without insight into its internal structure or code [4, 5]. Using this method, testers evaluate software based on input data and expected outputs, according to system specifications. The goal is to verify that the system responds correctly to inputs and generates expected outputs according to specifications [5, 6]. The tester plays the role of the end user and focuses on testing the input and output data, the behaviour of the application and its functional requirements. The advantage of this method is that it does not require technical knowledge about the code, it allows validation of the end user experience and can be used at any stage of software development. On the other hand, this method does not provide insight into the internal structure of the code and may miss implementation-level errors [6].

White-Box Testing involves an in-depth analysis of the internal logic and code structure of a software system [7]. A tester has insight into system implementation and uses technical skills to analyse paths, control flows, loops, and functions. This testing enables validation of not only functionality, but also performance optimization and identification of hidden bugs, improves quality of code through detailed analysis, and helps identify security vulnerabilities [5, 6]. Applying this method requires deep technical knowledge and access to code, and the software testing process can be very time-consuming.

Automated testing uses tools and scripts to execute predefined test cases, eliminating manual steps and speeding up the testing process [8]. It is used for regression testing, performance testing, scalability and load testing. Automated tools enable test repeatability and reduce the risk of human error, making them ideal for testing large and complex systems. Using this method significantly reduces the time and costs of testing, and enables quick detection of errors. This testing method requires a high initial investment in tools and employee training. Maintaining scripts can be complex if the software changes frequently.

Agile testing is an integral part of iterative and incremental development according to the Agile methodology [9]. Testing is done continuously throughout the development process, often in collaboration with developers and other team members. The focus is on quick feedback and adaptation to changes in user requirements. The advantages of this method are that it is flexible and easily adaptable to changes in user requirements [9]. Using this method improves communication between testers, developers, and users, which reduces the risk of major errors late in software development. Impeding factors for effective application of the method may be unclear priorities in software development or unclear user requirements.

3. METHODOLOGY

Two separate groups of researchers participated in the research. The first group consisted of experts from the fields of software development, information technology, software testing, management and economics. They were in charge of defining the evaluation criteria. Five key criteria were identified: efficiency, test coverage, ease of application, cost of implementation and reliability of results. Based on the technical requirements of the testing methods, the specifics of the projects and the available resources, the priorities of the criteria are defined in order to enable an objective assessment and the selection of the optimal testing method for different cases. Another group of researchers applied defined criteria for evaluating selected testing methods. Ratings are assigned based on practical experiences, user feedback and analysis of specific project needs. The PIPRECIA-S method was used to accurately determine the importance of each criterion. This method allows reliable ranking and selection of the optimal test method for a specific project.



3.1. PIPRECIA-S

The PIPRECIA-S method was used to determine the weighting coefficients of the criteria. Application of the PIPRECIA-S method implies that the importance of each criterion is compared with the importance of the first criterion. The main advantage of this method is its simplicity. Also, this method is very easily applied in group decision-making processes. The procedure for determining weight coefficients consists of five steps [3, 10].

The first step involves defining the evaluation criteria $C_j, j=1, \dots, n$ where n is the number of criteria taken into account when solving the problem. To determine the criteria, one can use the literature or the opinion of experts from the relevant field of research. The second step represents the determination of the relative importance of criteria s_j . It is necessary to determine the criterion (C_1) used as a basis for comparison. After that, starting from the second criterion, each criterion C_j is assigned the relative importance of criterion s_j based on Equation 1. Each criterion C_j is compared with the reference criterion C_1 .

$$s_j = \begin{cases} >1, & C_j > C_1 \\ =1, & C_j = C_1 \\ <1, & C_j < C_1 \end{cases}$$

Equation 1. Relative importance value s_j based on C_j and C_1

If criterion C_j is more important than criterion C_1 , it is assigned a value s_j that is greater than 1 and vice versa. If the criteria are equally important, then both criteria have an importance value of 1. The values s_j belong to the interval [0.6, 1.4]. The value of s_1 is always 1 and represents an estimate of the importance of the reference criterion C_1 .

In the third step, the value of the coefficient k_j is calculated based on Equation 2.

$$k_j = \begin{cases} 1, & j=1 \\ 2 - s_j, & j>1 \end{cases}$$

Equation 2. Calculation of the coefficient k_j

In the fourth step, the value of the coefficient q_j is calculated based on equation (3).

$$q_j = \begin{cases} 1, & j=1 \\ \frac{q_{j-1}}{k_j}, & j>1 \end{cases}$$

Equation 3. Calculation of the coefficient q_j

The last step involves the calculation of the relative weight of criteria w_j , where $0 \leq w_j \leq 1$ and $\sum_{k=1}^n w_k = 1$ based on Equation 4.

$$w_j = \frac{q_j}{\sum_{k=1}^n q_k}$$

Equation 4. Calculation of the relative weight of criteria w_j

The analysis was conducted on the basis of five key criteria: efficiency, test coverage, ease of application, cost of application and reliability of results.

Efficiency refers to the speed and ability of a method to detect and isolate software errors. Methods that enable rapid identification of critical issues are essential for projects with tight deadlines. The indicators used to evaluate this criterion are: the average error detection time and the number of detected errors in a certain period of time.

Test coverage measures how much system functionality or lines of code a method covers. High-coverage methods reduce the chance of undetected errors and enable detailed system analysis. The indicators used to evaluate this criterion are: percentage of functionality covered and percentage of code covered by the tests.

The criterion related to the simplicity of application evaluates how simple the method is to implement and use in real conditions. This includes the time required to train the team, the complexity of the tool, and the resources required for testing. The indicators used to evaluate this criterion are: the time required to learn the method and the required level of technical knowledge.

The cost of implementation includes the cost of tools, staff training, time spent on testing, as well as additional resources required for implementation. The indicators used to evaluate this criterion are: total costs per test cycle and maintenance costs of methods or tools.

Reliability of results refers to the accuracy and consistency of detected errors, as well as the method's ability to minimize false positive or negative results. The indicators used to evaluate this criterion are: the percentage of false positive and false negative results and the consistency of the results in different conditions.

3.2. RANKING SCALE

For each of the selected criteria, a ranking will be used based on the defined scale presented in Table 1. This scale follows the principles of the PIPRECIA-S method. The PIPRECIA-S method uses a specific range of values, typically 0.6 to 1.4, to determine the relative importance of criteria. Values less than 1.0 indicate reduced significance compared to the reference criterion, while values greater than 1.0 indicate increased significance. In order to enable the use of the known scale from 1 to 5 together with the PIPRECIA-S principles, a conversion was made between these scales.



3.3. PRIORITIZATION OF CRITERIA

The ranking of criteria according to their importance in the evaluation process of software testing methods is shown in Table 2. At the beginning of the research, the criteria were ranked on a scale from 1 to 5. Grade 1 indicates the lowest priority, and grade 5 the highest. The obtained scores were aggregated, and the average values were used as input data for the PIPRECIA-S method. Through an iterative process, participants adjusted their ratings. The criteria are ranked for the purposes of this research, but these results may differ depending on the software development context, requirements and goals of the specific project. Table 2 also shows the evaluation results of software testing methods according to defined criteria.

4. RESULTS

The relative importance of software testing methods in relation to the Efficiency criterion is shown in Table 3. The evaluations of software testing methods based on this criterion were made based on the following facts:

- Black Box achieves a score of 5 because it enables quick and efficient detection of errors from the perspective of the end user, without the need to know the internal structure of the system;
- White-Box gets a score of 4 because it provides detailed code analysis and a high level of test coverage, but requires significant time and technical knowledge;
- Automated testing achieves a score of 5 because it significantly speeds up the testing process, enables repeatability and covers a large number of scenarios in a short time and

Table 1. Ranking scale

Description	Rank	PIPRECIA-S scale
Very bad	1	0.60
Bad	2	0.80
Satisfactory	3	1.00
Good	4	1.20
Excellent	5	1.40

Table 2. Evaluation results of software testing methods according to defined criteria

Specification	Notation	Rank	Black Box	White Box	Automated	Agile
Efficiency	C_1	5	Excellent (5)	Good (4)	Excellent (5)	Good (4)
Testing coverage	C_2	4	Satisfactory (3)	Good (4)	Excellent (5)	Satisfactory (3)
Ease of application	C_3	3	Good (4)	Satisfactory (3)	Satisfactory (3)	Excellent (5)
Application price	C_4	2	Excellent (5)	Satisfactory (3)	Satisfactory (3)	Good (4)
Reliability of results	C_5	5	Good (4)	Excellent (5)	Excellent (5)	Good (4)

Table 3. Relative importance of the considered software testing methods in terms of criterion C_1 - Efficiency

		s_j	k_j	q_j	w_j
A_1	Black Box Testing		1	1	0,19
A_2	White Box Testing	1,2	0,80	1,25	0,24
A_3	Automated Testing	1,4	0,60	1,67	0,32
A_4	Agile Testing	1,2	0,80	1,25	0,24
				5,17	1,00



- Agile testing gets a score of 4 because it effectively supports iterative development and enables rapid adaptation to changes, but may be less focused on deeper analysis of specific problems.

The relative importance of software testing methods in relation to the Test Coverage criterion is shown in Table 4. The evaluations of software testing methods based on this criterion were made based on the following facts:

- Black-Box gets a score of 3 because it covers functionality from the user's perspective, but does not include internal logic errors or system structure;
- White-Box achieves a score of 4 because it enables deeper testing of internal system components and detailed code analysis;
- automated testing gets a score of 5 because it enables wide coverage by testing different scenarios and a large volume of tests with minimal human effort and
- Agile testing gets a score of 3 because it focuses on continuous adaptation and iteration, but does not provide comprehensive coverage of all scenarios.

Table 5 presents the relative importance of software testing methods in relation to the Ease-of-Use criterion, with the following conclusions:

- Black-Box gets a score of 4 because it is easy to implement and does not require technical knowledge about the system structure;

- White-Box achieves a score of 3 because it requires detailed knowledge of the code and more technical skills, which makes it difficult to apply in some situations;
- Automated testing gets a score of 3 because it requires initial setup of scripts and tools, which can be technically demanding and
- Agile testing achieves a score of 5 because it enables quick adaptation and easy integration into iterative software development processes.

The relative importance of software testing methods in relation to the Cost of Implementation criterion is shown in Table 6. The evaluations of software testing methods based on this criterion were made on the basis of the following conclusions:

- Black-Box achieves a score of 5 because it is cost-effective and requires minimal resources compared to other methods;
- White-Box gets a score of 3 because it requires more time, professional staff and resources, which increases the cost of implementation;
- Automated testing gets a score of 3 because the initial setup of tools and scripts can be expensive, but the costs are reduced in the long run and
- Agile testing gets a score of 4 because the costs are adjusted to the needs of the project and are usually lower compared to methods that require complex tools.

Table 4. Relative importance of the considered software testing methods in terms of criterion C_2 - Testing coverage

		s_j	k_j	q_j	w_j
A_1	Black Box Testing		1	1	0,20
A_2	White Box Testing	1,2	0,80	1,25	0,25
A_3	Automated Testing	1,4	0,60	1,67	0,34
A_4	Agile Testing	1,0	1,00	1,00	0,20
				4,92	1,00

Table 5. Relative importance of the considered software testing methods in terms of criterion C_3 - Simplicity of application

		s_j	k_j	q_j	w_j
A_1	Black Box Testing		1	1	0,21
A_2	White Box Testing	1,0	1,00	1,00	0,21
A_3	Automated Testing	1,0	1,00	1,00	0,21
A_4	Agile Testing	1,4	0,60	1,67	0,36
				4,67	1,00



Table 7 shows the relative importance of software testing methods in relation to the criterion Reliability of results, with the following conclusions:

- Black-Box receives a score of 4 because the results reliably reflect functional errors, but do not include internal system problems;
- White-Box achieves a score of 5 because it provides detailed and precise information about errors within the system;
- Automated testing receives a score of 5 because it provides consistent and repeatable results with minimal human error and
- Agile testing gets a score of 4 because it provides reliable results in the context of iterative cycles but may miss details due to the fast pace of development.

Figure 1 shows the final ranking order of the analysed software testing methods based on the importance of the given criteria.

5. DISCUSSION

The evaluation results of the four selected software testing methods based on the five analysed criteria show significant differences in the performance of the analysed methods.

Automated testing received the highest overall value and was ranked first. This suggests that automated testing provides the best results compared to other methods, primarily due to its high efficiency, test coverage and reliability of results.

Table 6. Relative importance of the considered software testing methods in terms of criterion C_4 - Cost of application

		s_j	k_j	q_j	w_j
A_1	Black Box Testing		1	1	0,24
A_2	White Box Testing	1,0	1,00	1,00	0,24
A_3	Automated Testing	1,0	1,00	1,00	0,24
A_4	Agile Testing	1,2	0,80	1,25	0,29
				4,25	1,00

Table 7. Relative importance of the considered software testing methods in terms of criterion C_5 - Reliability of results

		s_j	k_j	q_j	w_j
A_1	Black Box Testing		1	1	0,18
A_2	White Box Testing	1,4	0,60	1,67	0,30
A_3	Automated Testing	1,4	0,60	1,67	0,30
A_4	Agile Testing	1,2	0,80	1,25	0,22
				5,58	1,00

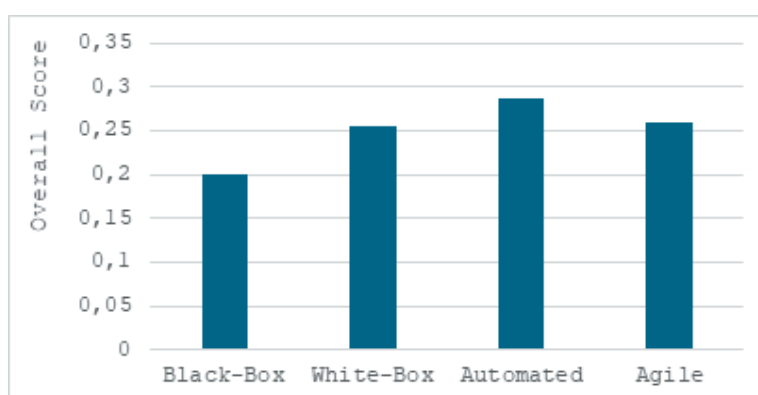


Figure 1. Final ranking of software testing methods



These factors make this method particularly suitable for situations that require high precision and coverage in testing, although the cost of implementation is relatively lower compared to the advantages that this method provides.

Agile testing, which ranked second, shows balanced performance across all criteria. Its strongest point is its ease of deployment, making it an ideal choice in agile environments where flexibility and speed are key. However, its reliability of results lags behind automated testing, which can be a challenge in complex projects.

White-Box Testing took third place with overall value. Its advantages are the efficiency and reliability of the results, but the relatively lower score in test coverage and ease of application indicates that this method is more suitable for specific cases where a detailed understanding of the internal logic of the system is required.

Black-Box Testing ranked last. Although this method has the advantage of simplicity of application, its results in criteria such as reliability of results and efficiency show that it is less suitable for projects where reliability and coverage are crucial. This indicates that this method is best suited for the initial stages of testing or for less complex systems.

Although agile testing and white-box testing have achieved similar results, the choice between the two methods depends on the specific requirements of the project. Automated testing stands out as the most effective option, while Black-Box Testing remains the least effective choice for complex projects.

6. CONCLUSION

The analysis carried out using the PIPRECIA-S method showed that different software testing methods have specific advantages and limitations depending on the analysed criteria. Automated testing has been identified as the most reliable and efficient method, thanks to its ability to provide high test coverage and reliable results, with relatively low implementation costs. On the other hand, Black-Box Testing, while simple to implement, falls short in key categories, making it more suitable for less complex projects. Agile testing has proven to be a method that balances flexibility and reliability, while White-Box Testing provides high reliability in specific cases, but requires greater technical knowledge. Based on the obtained results, the choice of the optimal testing method should be adapted to the specific requirements of the project, taking into account the criteria of efficiency, costs and complexity of implementation.

Conducting this research has practical and scientific significance. On the one hand, the results can help software teams and managers to make better decisions when choosing testing methods. On the other hand, the research contributes to the theoretical understanding of the application of multicriteria analysis in the domain of software testing. Future research can extend this analysis by including additional methods and criteria, to enable even more precise selection of test strategies in different industrial settings.

REFERENCES

- [1] M. Bajjouk, M. E. Rana, C. Ramachandiran and S. Chelliah, "Software testing for reliability and quality improvement", *Journal of Applied Technology and Innovation*, vol. 5, no. 2, 2021, pp. 40-46. [Online]. Available: https://www.researchgate.net/publication/359617048_Software_testing_for_reliability_and_quality_improvement [Accessed: February 10, 2025].
- [2] U. Azeem & A. Abhineet, "Importance of Software Testing in the Process of Software Development", *International Journal for Scientific Research & Development*, vol.6, no.12, 2019, pp. 141-145. [Online]. Available: https://www.researchgate.net/publication/331223692_Importance_of_Software_Testing_in_the_Process_of_Software_Development [Accessed: January 22, 2025].
- [3] S. Popović, D. Viduka, A. Bašić, V. Dimić, D. Djukic, V. Nikolić & A. Stokić, "Optimization of Artificial Intelligence Algorithm Selection: PIPRECIA-S Model and Multi-Criteria Analysis", *Electronics*, vol. 14, no. 3, 2025, pp. 562. [Online]. Available: <https://doi.org/10.3390/electronics14030562> [Accessed: March 11, 2025].
- [4] U. Mubarak Albarka, "A Study of Software Testing: Categories, Levels, Techniques, and Types", *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 5, no. 6, 2019, pp. 32-40. [Online]. Available: <http://dx.doi.org/10.36227/techrxiv.12578714> [Accessed: March 10, 2025].
- [5] S. Nidhra & J. Dondeti, "Black Box and White Box Testing Techniques - A Literature Review", *International Journal of Embedded Systems and Applications*, vol. 2, no. 2, 2012, pp. 29-50. [Online]. Available: <http://dx.doi.org/10.5121/ijesa.2012.2204> [Accessed: March 10, 2025].
- [6] T. Capote, "A comparative study of Black box and White box testing techniques in modern software development", *Frontiers In Engineering And Technology (FET)*, vol. 5, no. 1, 2024, pp.1-7. [Online]. Available: <https://iaeme.com/Home/issue/FET?Volume=5&Issue=1> [Accessed: March 5, 2025].



- [7] M. Ehmer & F. Khan, "A Comparative Study of White Box, Black Box and Grey Box Testing Techniques", *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 6, 2012, pp.12-15. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2012.030603> [Accessed: March 10, 2025].
- [8] P. Mahajan, H. Shedge, & U. Patkar, "Automation Testing In Software Organization", *International Journal of Computer Applications Technology and Research*, vol. 5, no. 4, 2016, pp. 198-201. [Online]. Available: <http://dx.doi.org/10.7753/IJC-ATR0504.1004> [Accessed: March 15, 2025].
- [9] W. Gopalkrishna, J. Shah & P. Nanavati, "Agile software development", *International Journal of Engineering Applied Sciences and Technology*, vol. 5, no. 12, 2021, pp.73-78. [Online]. Available: <http://dx.doi.org/10.33564/IJEAST.2021.v05i12.011> [Accessed: March 18, 2025].
- [10] D. Stanujkic, D. Karabasevic, G. Popovic & C. Sava, "Simplified Pivot Pairwise Relative Criteria Importance Assessment (Piprecia-S) Method" , *Romanian Journal of Economic Forecasting*, vol. 24, no. 4, 2021, pp. 141–154. [Online]. Available: https://www.researchgate.net/publication/357736315_SIMPLIFIED_PIVOT_PAIRWISE_RELATIVE_CRITERIA_IMPORTANCE_ASSESSMENT_PIPRECIA-S_METHOD [Accessed: December 20, 2024].



DYNAMICS OF MODERN WEB SCRAPING: CLIENT EXPECTATIONS VS TECHNICAL REALITY

Umar Khalid*,
[0000-0002-3693-9859]

Armeen Shahid
[0009-0002-9790-8914]

Scraping Solution Ltd.,
London, United Kingdom

Abstract:

The emergence of the World Wide Web has increased the amount of information that is easily accessible on the Internet. This information could be copied manually or scraped by using automated tools. Web scraping also known as web mining, data extraction, or web harvesting refers to extracting raw data from different websites and converting it into useful information. This information can be used by researchers, business organizations, social scientists, and health care professionals for data analysis, understanding the specific market, finding potential clients, people, or market sentiments, or health care wide spectrum studies. This paper focuses on an overview of the changing dynamics of web scraping, a never-ending marathon of web scraping developers and anti-scraping service providers later this article discusses the lack of understanding among clients who still view web scraping as merely a simple process automation to copy data, overlooks the growing complexities involved in navigating the evolving landscape of AI-driven websites and anti-scraping measures.

Keywords:

Web Scraping, Data Extraction, Client Expectations, Automation, Anti-Bot Systems, Technical Challenges.

INTRODUCTION

Data is an essential part of any research, whether it be academic, marketing, or scientific. It is essential for businesses and organizations as it assists their decision making and especially, currently, most of the data can be found on the internet and is available publicly all around the globe on websites [1]. As the internet becomes an ever-growing repository of information, the need to efficiently gather, process, and analyze this data has grown exponentially. Businesses and organizations often face several challenges in acquiring high- quality and required information. Early web scraping techniques involved basic techniques, such as downloading web pages and using regular expressions or simple scripts to extract data. It was often a manual, semi-automated, or occasionally fully automated process, lacking the sophistication and automation of today. Also, it is difficult to access unstructured data and dynamic content generated by JavaScript or embedded within complex HTML structures.

Correspondence:

Umar Khalid

e-mail:

umarkhalid@scrapingsolution.com





Moreover, there were few, if any, legal restrictions or formal policies in place. Bot-blocking or anti-scraping systems were virtually non-existent, and at most, basic IP blockers were used—easily bypassed with minimal effort. Today, traditional data acquisition techniques are not efficient and an unpleasant procedure for real-time data collection is not flexible enough for large amounts of data [2].

Modern web scraping involves automatically extracting large bulks of data from websites using software tools and scripts. Data could be in any form such as text, images, videos, or other forms. Throughout the last decade web scraping has evolved to an entirely new level. With the advancement in technology. It has become more powerful, automated, versatile, flexible, and reliable. The business and research sectors depend heavily on the quality, accuracy, and comprehensiveness of data. Consequently, client expectations are far beyond the true concepts of web scraping. Their expectations are not aligned with reality or are still trapped in the previous decade. As a result, web scrapers frequently encounter challenges in delivering the required output. These issues stem from the inherently complex nature of web scraping along with dynamic websites, and anti-scraping measures like CAPTCHAs, IP-blocking, rate limiting, and evolving legal constraints around data privacy and intellectual property.

The main objectives of this paper are to shed light on the often-overlooked ground realities of web scraping, a technically strong process but its challenges may impact its effectiveness and efficiency. On the other hand, the paper also delves into the expectations that clients typically have when commissioning web scraping projects.

2. OVERVIEW OF WEB SCRAPING

Web scraping also known as web extraction and web harvesting is a technique to extract both structured and unstructured data from websites and transform it into an organized form that can be stored in databases, or comma-separated values [3]. Web scraping is the method involved with separating or extracting information off the web programmatically and changing it into an organized dataset. It is particularly important in fields such as Business Intelligence, Artificial Intelligence, Data Science, Big Data, Cyber Security, Digital Marketing, Sentiment Analysis, and E-commerce development/management in the modern age. Types of scraping discussed by authors in [4] are: Extracting data with an HTML parser or regular expression matching,

the second one is by using an application programming interface referred to as API. Web scraping techniques are broadly used in web indexing, web mining, web data integration and data mining [5].

Website investigation, website crawling, and information getting sorted out are the three essential cycles of internet scraping [6]. Due to the wide number of open devices and libraries that offer productive executions of a significant part of the necessary usefulness, web scraping is a straightforward and pretty simple process in general.

In today's era of big data, the demand for relevant information has driven individuals and companies to extensively use web scraping as a crucial tool for gathering data. Companies that aggregate product availability and offer price comparisons have built billion-dollar businesses on the foundation of web scraping. Similarly, they employ web scraping to provide consumers with valuable insights, highlighting the significant role of web scraping in modern business for enabling competitive analysis, market research, and customer decision-making [7]. While others might do web scraping for their interests such as job hunting or getting housing offers [8].

The current scenario of web scraping in the generative AI era is quite blurry but at the same time fast-paced, where companies like Cloudflare, Datadome and others are heavily investing in the development of anti-scraping systems with zero tolerance of allowing any kind of scraper or bot to take any data from webpage, however on the other side, big startups like ScrapingBee, Zyte api and Oxylabs have been started in recent years to develop web unblockers and proxy Api's to bypass the system developed by Cloudflare or Datadome, Both continuously working to overcome each other's and selling their solution to web developers and scraping developers, a rate race is still ongoing and their market share increasing so the cost of web scraping and data mining. Therefore, with every effort made, data theft news from highly protected systems hit the news every other day [9] and the legal battles are also part of the news now and then [10].

3. DESIGN OF WEB SCRAPING BOTS

A typical design of a bot comprises three phases: website analysis, web scraping, and data organization. The first is the most critical and challenging while the last one is quite straightforward and repetitive. Even with AI advancements, the initial phase still demands significant human involvement, whereas the subsequent stages are comparatively easier, more automated, and can be AI-assisted.



3.1. WEB ANALYSIS

Basic analysis examines the channel through which data is being transferred between the website and the server. Mostly its either get or post request respond by an HTML, XML, or JSON response. The interaction can be either directly with the website server/database or through an API.

Lately, Modern web analysis also includes a thorough study of the anti-bot services deployed on the website along with its version and the study of recent version updates. This analysis further helps the developer to either develop a function to bypass this blocking service itself or use a relevant third-party service.

3.2. WEB CRAWLING

The phase of developing and running a script or program that automatically browses the website and retrieves the required data. It is comprised of languages such as Python, and R. Web scraping libraries; request, beautiful soup, scrapy, and selenium.

3.3. DATA ORGANIZATION

The collected data need to be stored in an organized manner along with the need for data cleaning and pre-processing. Many programming languages, such as R and Python, contain Natural Language Processing (NLP) libraries and data manipulation functions that are useful for cleaning and organizing data [11].

Web scraping could be done in several languages each offering its unique tools and libraries for facilitating the process [12]. Approaches using Python packages discussed in [13] are Regular Expressions, BeautifulSoup, Scrapy, and Lxml. Along with popular headless browsers, selenium is efficient to be used.

4. CLIENT EXPECTATIONS AND TECHNICAL REALITIES OF WEB SCRAPING

This section focuses on the challenges faced by web scraping developers or scraping service providers, highlighting the issues regarding client expectations and the reality of web scraping along with the technical challenges in web scraping.

The client expects and demands something that exactly matches the requirement or idea they are after, sometimes along with non-understanding or rigid behavior particularly in terms of data elements, com-

prehensiveness, and speed without even knowing the technical difficulties involved. Therefore, the scrapers encounter various challenges such as finding a reliable source for the data (if not mentioned by the client) reflecting the exact information asked, website structure changes, dynamic content, CAPTCHA, and anti-bot measures that may include IP-rate limiting, and device fingerprint monitoring. Although Web scrapers are determined to deal with anti-scraping measures that focus on protecting the content and preventing unauthorized access to large datasets these anti-bot services always keep on updating and sometimes their way to detect a bot may change every day which starts a rat race of analyzing and finding the solution to bypass these services. These challenges may require a lot of human, financial, and computational resources, and could lead to delays in delivery and impact client expectations.

From the clients' perspective, web scraping is considered a low-cost procedure. However, the cost may escalate with the need for advanced scraping techniques, maintenance, and overcoming the ever-evolving anti-scraping services. Scaling a scraping operation also involves significant technical challenges, such as managing server loads, handling a vast number of requests, and dealing with sophisticated anti-scraping measures. While the client may expect the solution to scale effortlessly as their data needs to grow, the complexities often require additional resources and expertise.

On legal grounds, clients may assume that all web scraping activities are lawful and require no specific permissions. However, considering the legalities, scraping large volumes of web data involves serious challenges, including compliance with data privacy laws, intellectual property regulations, and the terms of service set by websites. Both service providers and clients must ensure that scraping operations are conducted within the bounds of applicable legal frameworks to avoid potential liabilities. While there has been an increase in tools and advancement in technologies that can be used for Web Scraping legality and ethics of data collection from the Web are still "grey areas" [11]. Many websites now also require users to agree to "terms and conditions" that explicitly prohibit data scraping, which increases the challenges of web scrapers [12].

In the following section, I will present two case studies from projects undertaken by Scraping Solution. These examples highlight real-world scenarios where web scraping efforts either struggled, escalated in complexity, or were ultimately cancelled. The case studies reflect the dynamic nature of web scraping, where techni-



cal complexities such as anti-bot mechanisms, evolving site structures and client misconceptions significantly impacted project outcomes. The reason for including these specific cases, where the projects faced significant challenges, is because they offer more impactful insights and help steer the article toward a conclusive understanding of the realities surrounding web scraping.

5. CASE STUDY 1: WEB SCRAPING FOR YAHOO FINANCE

5.1. CLIENT REQUIREMENTS

The client approached Scraping Solution to develop a custom scraper for Yahoo Finance, specifically to scrape articles from the news section on a regular basis. Initial discussions revealed the client's desire for a robust, high-speed scraping solution with minimal technical requirements. They expressed an understanding of the web scraping process, which was considered during client profiling. Client profiling at Scraping Solution is a standardized procedure that involves assessing the client's background, previous project ratings, response rate, and other metrics from freelance platforms.

5.2. TECHNICAL INVESTIGATION

Upon initiating the project, Scraping Solution's established methodology involves checking four key aspects:

1. Connection with the Server:

Yahoo Finance did not have an available API for its news section. Thus, it became evident that a browser automation tool like Selenium or a request-based solution would be needed.

2. Static vs. Dynamic Content:

Yahoo Finance's response was identified as dynamic, meaning the HTML structure varied with each request, potentially altering with geographic location. This led to the realization that the scraper would need to handle dynamic HTML responses and geographic fingerprinting, requiring country-specific IP rotation or VPN usage to maintain consistency.

3. Anti-Bot Protection:

The website employed anti-bot mechanisms, though they were not overly restrictive. A simple IP rotation or VPN setup would suffice for maintaining access to Yahoo Finance's articles without triggering anti-bot blocks.

4. Data Storage Format:

As the project involved scraping article data, the data formatting requirement was relatively simple compared to more complex product scraping tasks.

5.3. CLIENT FEEDBACK AND PROJECT OUTCOME

The research was presented to the client along with a recommended solution that employed Selenium for browser automation, supported by IP rotation or VPNs to mitigate the impact of geographical variation in responses. Despite the technical depth and the low-cost recommendation for VPN or IP rotation services, the client rejected the proposed solution, stating that they were unwilling to use VPNs or browser automation tools.

This led to a significant gap between client expectations and technical realities. The client requested a solution that would work without fingerprint rotation or browser automation—an approach that was technically impossible given the nature of Yahoo Finance's security measures and dynamic content delivery. Despite professional communication and attempts to clarify these complexities, the project ended in dispute and was ultimately cancelled, with extensive unpaid research work already completed.

6. CASE STUDY 2: PRODUCT SCRAPING FOR SHOPIFY

6.1. CLIENT REQUIREMENT

A client engaged Scraping Solution to automate the scraping of product data from their supplier's website and format it into a Shopify-uploadable template. The client also requested a cron job to regularly update new products and price changes twice a week. The goal was complete automation, allowing the client to focus solely on the data without needing manual intervention.

6.2. TECHNICAL INVESTIGATION

The website was initially straightforward, with no significant anti-bot services in place. However, the primary challenge lay in scraping product variants, which varied dynamically across different product categories. Some products featured variants like size, color, or weight, and their prices were embedded in JavaScript, requiring additional interaction (click events) to retrieve the pricing data.



The proposed solution was a hybrid of techniques

1. API and Requests:

The scraper would combine API requests, traditional GET/POST methods, and Selenium automation to handle the dynamic JavaScript-based variant data.

2. Handling Product Variants:

A significant portion of the effort was spent coding around the various product departments, each of which required different techniques for parsing product variants. This involved interacting with dynamic elements on the webpage to extract accurate pricing.

3. Automated Updates:

It was recommended that a Flask API or cron job be implemented to check for updates twice a week. Simple IP rotation was also suggested to avoid potential scraping blocks, ensuring the system remained reliable in the long term.

4. Data Formatting:

The scraped data was to be parsed and formatted into a Shopify-uploadable template. This involved heavy data parsing and ensuring compatibility with Shopify's import format.

5. Ongoing Maintenance:

Scraping Solution also proposed hosting the system on their own servers, ensuring real-time monitoring and quick fixes in case of website changes or system crashes. Regular updates to the script were anticipated, as scraping scripts typically require monthly adjustments based on experience.

6.3. CLIENT FEEDBACK AND PROJECT OUTCOME

The client appreciated the research and detailed solution but ultimately decided not to move forward with the development. The client's company was unable to allocate the necessary budget, even though the ongoing maintenance costs were under \$200 per month. This hesitation was attributed to two common misconceptions: the belief that web scraping is a simple copy-paste automation process, and a reluctance to incur ongoing monthly expenses for a task they felt could be handled by an in-house hire.

7. DISCUSSION

The complexities and challenges highlighted in both case studies reflect a larger pattern in the web scraping industry—misconceptions about the technical depth required for effective and scalable scraping solutions. As demonstrated by the clients in the examples, a recurring issue is the reluctance to adopt necessary tools, technologies, and ongoing support systems, stemming from a lack of understanding of the nuances involved in web scraping.

In the first case study, the client's hesitation to utilize essential technologies such as browser automation tools and IP rotation reveals a fundamental gap in comprehension. Web scraping is not simply a copy-paste automation task; it is an intricate process that requires dealing with dynamic content, anti-bot systems and geographical restrictions. The assumption that data can be scraped without addressing these complexities reflects a prevalent issue where clients expect scraping solutions to operate with minimal intervention. Browser automation and IP rotation are often indispensable in bypassing modern anti-bot protections and content variations and any attempt to avoid these tools severely limits the scraper's capability. However, such tools are sometimes viewed as unnecessary add-ons, even though they are critical for achieving the desired outcome.

Similarly in the second case study, the client viewed web scraping as a low-cost process assuming that once the system was developed, it would require little to no ongoing investment. This is another common misunderstanding where clients do not fully grasp that scraping environments are not static. Websites frequently undergo changes in their structure, security measures or data presentation which can cause a scraping script to break. The assumption that scraping requires no continuous input from the developer overlooks the necessity of regular updates, error handling and maintenance. In this case, even though the proposed monthly cost was quite low, the client could not just apprehend the expense as he was viewing the scraping process as a "set it and forget it" operation.

This tendency among SMEs and sole traders to resist ongoing costs for web scraping maintenance often stems from a traditional mindset where automation is expected to function indefinitely without requiring human intervention. In reality, automation in the context of web scraping requires constant adaptation, especially when it comes to dynamic websites, complex data structures and anti-scraping measures such as CAPTCHAs or JavaScript-based protection systems.



The reluctance to invest in regular updates and maintenance may also be tied to a perception that the ongoing costs are unjustified or even exploitative. Small business owners in particular, can view monthly charges as “paying for nothing” or mistakenly believing that once the code is written, the process will run indefinitely without further input from the developer. This notion deduces discounts the technical expertise required to ensure that the scraper continues to function smoothly, avoiding crashes, detecting changes in the target website and ensuring data integrity over time.

Both cases exemplify how the lack of understanding of the technical particulars of web scraping can create conflicts between developers and clients. Whether it is resistance to adopt necessary tools like IP rotation or browser automation or the refusal to accept the need for ongoing maintenance. These issues stem from a failure to recognize that web scraping is far more complex than simple automation. As long as these misconceptions persist, developers will continue to face challenges in bridging the gap between client expectations and the technical realities of web scraping.

Ultimately, web scraping is not just about extracting data, but it is about building a robust, adaptable system that can respond to changes in the environment while continuing to deliver value. And that requires more than a one-time investment; it demands ongoing collaboration and support.

8. CONCLUSION AND FUTURE DYNAMICS

Web scraping has become an essential tool for businesses seeking data-driven insights, maintaining their eCommerce stores or managing their marketing strategies, whether for one- off projects or long-term automation solutions. While many clients may approach scraping as a simple, one-time task but the reality often proves more complex. Even in short-term projects, factors such as dynamic content, anti-scraping measures and evolving web technologies mean that the process requires a more thoughtful and flexible approach. For clients, understanding the nuances of web scraping including the potential need for ongoing support or updates can ensure smoother project execution and better long-term outcomes.

Looking toward the future, it's clear that web scraping will become more challenging. As both bot blockers and scrapers continue to invest massively in human and capital resources, the battle over data access will only grow more complex. With websites adopting increasingly sophisticated anti-scraping techniques, web scraping businesses will need more advanced solutions to extract the data they rely on for AI, eCommerce, marketing and

analytics. Although scraping is becoming more technically complex and expensive, it will still remain indispensable in a world that thrives on data-driven decision-making.

REFERENCES

- [1] I. S. Almaqbali, "Web scrapping: Data extraction from websites," *J. Student Res.*, 2019.
- [2] R. Vording, "Harvesting unstructured data in heterogeneous business environments; exploring modern web scraping technologies," M.S. thesis, Univ. Twente, 2021.
- [3] K. Mehta, M. Salvi, R. Dand, V. C. Makharia, and P. Natu, "A comparative study of various approaches to adaptive web scraping," in *Lecture Notes in Electrical Engineering*, vol. 601, Springer, 2020, pp. 1245–1256. doi: 10.1007/978-981-15-1420-3_136.
- [4] M. Dogucu and M. Çetinkaya-Rundel, "Web scraping in the statistics and data science curriculum: Challenges and opportunities," *J. Stat. Data Sci. Educ.*, vol. 29, no. S1, pp. S112–S122, 2021, doi: 10.1080/10691898.2020.1787116.
- [5] M. S. Parvez, K. S. A. Tasneem, S. S. Rajendra, and K. R. Bodke, "Analysis of different web data extraction techniques," in *Proc. 2018 Int. Conf. Smart City Emerging Technol. (ICSCET)*, Mumbai, India, Jan. 2018, pp. 1–5, doi: 10.1109/ICSCET.2018.8537333.
- [6] P. H. Milev, "An assessment of information systems for the indexation and analysis of online publications," *Spisanie "Biznes Upravljenje"*, 2017.
- [7] Y. Neil, *Web Scraping the Easy Way*, Georgia Southern Univ., 2016.
- [8] G. Pérez Molano, "An overview of web scraping: Technical aspects and exercises," *PRC Repository*, vol. 8, 2023.
- [9] Reuters, "Multiple AI companies bypassing web standard to scrape publisher sites, licensing firm says," *Reuters*, 2024.
- [10] T. H. Bureau, "Google sued in the U.S. over data scraping for AI," *The Hindu Bureau*, 2023.
- [11] V. Krotov, L. Johnson, and L. Silva, "Tutorial: Legality and ethics of web scraping," *Commun. Assoc. Inf. Syst.*, vol. 47, no. 1, pp. 22, 2020, doi: 10.17705/1CAIS.04724.
- [12] A. Luscombe, K. Dick, and K. Walby, "Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences," *Qual. Quant.*, vol. 56, no. 3, pp. 1023–1044, 2022, doi: 10.1007/s11135-021-01164-0.
- [13] I. Valova, T. Mladenova, G. Kanev, and T. C. Halacheva, "Web scraping - state of the art, techniques and approaches," in *Proc. 31st National Conference with International Participation (TELECOM)*, Svishtov, Bulgaria, Oct. 2023, pp. 1–4, doi: 10.1109/TELECOM59629.2023.10409723.



PERFORMANCE ANALYSIS OF HAAR CASCADE-BASED FACE DETECTION IN MULTI-FACE IMAGES UNDER DIVERSE COMPRESSION ALGORITHMS

Ivan Šarkoćević^{1*},
[0009-0005-3387-442X]

Vladimir Maksimović²,
[0000-0001-8479-0027]

Branimir Jakšić²,
[0000-0001-6683-0021]

Petar Spalević³,
[0000-0002-6867-7259]

Đoko Bandur²
[0000-0001-9034-6854]

¹Kosovo and Metohija Academy
of Applied Studies,
Zvečan, Serbia

²Faculty of Technical Sciences,
University of Pristina in Kosovska
Mitrovica,
Kosovska Mitrovica, Serbia

³Singidunum University,
Belgrade, Serbia

Correspondence:

Ivan Šarkoćević

e-mail:

ivan.sarkocevic@akademijakm.edu.rs

Abstract:

With the tremendous development of face detection systems, there is a growing need to achieve highly accurate detection results in images compressed using different compression algorithms. This manuscript provides face detection analysis in images representing different numbers of faces (1, 3, 5, and 10 faces) from both frontal and non-frontal perspectives. The images extracted from the FDD (Face Detection Dataset) database were compressed using three different compression algorithms - JPEG, JPEG2000, and SPIHT, for different bits-per-pixel values.

The analysis was performed by using the Haar Cascade Classifier, implemented in Python. The quality of face detections was determined using the objective measures: F-measure (based on reference values from the GroundTruth images) and Det.F (number of detected faces). Based on the results presented in the tables, it can be concluded that face detection behaves slightly differently depending on the value of bits-per-pixel and the applied compression algorithm, but vastly differently depending on the angle of perspective.

The Haar Cascade Classifier has proven to be the best solution when it is necessary to perform face detection in compressed, frontal face images, especially for a small number of faces; non-frontal images with a large number of faces have proven to be the most challenging assignment for the Haar Cascade Classifier.

Keywords:

Face Detection, Compression, Image Processing, Haar Cascade Classifier, F-measure.

INTRODUCTION

Regarding modern-day multimedia systems, it is nearly impossible to find a system that doesn't utilize image, video, or audio compression. The advancements in technology have led to a growing need for image processing - from everyday applications to critical professional fields, such as medical image analysis, sensor networks, security systems, and television [1].

Cameras capture images in analog format; the images are then converted into digital formats for efficient processing, transmission, and storage [2]. Uncompressed images demand substantial storage capacity and computational resources for processing, in addition to greater bandwidth for transmission across communication networks.



As a result of the presence of these challenges, image processing and compression techniques have become a substantial focus of research in digital signal processing [1].

The aim of developing image compression techniques is to remove redundant or unnecessary data. The techniques are categorized based on their ability to recreate the original image from the compressed data. There are two types of digital image compression: lossy compression and lossless compression. Lossy compression allows for greater compression ratios by selectively discarding redundant data. Lossless compression keeps all the original image data intact, thereby allowing for perfect reconstruction without any loss of information [2] [3]. The choice of compression methods, techniques, and algorithms is based on the requirements of the application. A variety of techniques have gained prominence in digital image processing. JPEG (Joint Photographic Experts Group), JPEG2000, and SPIHT (Set Partitioning in Hierarchical Trees) have appeared to be three of the most prevalent standards, each offering distinct advantages [1].

The JPEG compression standard has become well-known and widely used in multiple areas of application, including digital photography, web content, and medical diagnostics. The JPEG uses the Discrete Cosine Transform (DCT), so that it can divide an image into different frequency components. This process separates important visual details from less noticeable ones. The image becomes easier to compress effectively [4] [5].

The JPEG2000 compression standard was developed as a successor to the original JPEG standard to address the demands of emerging applications. It delivers advanced functionality while achieving superior compression performance [6] [7]. Unlike traditional JPEG standard, which uses DCT, as mentioned above, the newer JPEG2000 compression standard uses the Discrete Wavelet Transform (DWT) as its mathematical framework [1].

While JPEG is a lossy compression method [8], and JPEG2000 offers both lossy and lossless compression [7], the SPIHT algorithm is entirely lossless [8]. SPIHT represents a wavelet-based compression method that is very efficient and highly effective. It outperforms the majority of traditional techniques by achieving superior image quality at equivalent compression ratios. The quality is measured by the Peak Signal-To-Noise Ratio - PSNR. The Embedded Zerotree Wavelet (EZW) algorithm represents the mathematical framework for the SPIHT coder [8].

The growing intelligence of computer systems has had a tremendous impact on the ongoing advancement of technology, starting a new era of human-computer interaction. An important example of this interaction is face detection - a special case of object detection. Object detection is used to identify objects in digital images and videos, such as human faces or vehicles. This characteristic of object detection leads to an extensive application of the technique, including face recognition, pedestrian detection, and surveillance systems [9].

Face detection is the critical first step that enables subsequent stages of face analysis, such as face recognition and verification. This stage identifies and locates human faces in a digital image so that further analysis can be performed. The detection process consists of analyzing input images through pattern differentiation, isolating facial features from background elements and outputting spatial coordinates for each detected face [9].

The authors in [10] examined the effect of compression on face detection using Haar-like features, while the authors in [11] examined frontal and non-frontal face detection using deep neural networks. The aim of this paper is to examine how JPEG, JPEG200, and SPIHT compression techniques affect face detection using the Haar Cascade Classifier, considering both frontal and non-frontal images.

The remainder of the paper is structured as follows: Section 2 outlines the system model which provides the fundamental setting for a detailed analysis of the face detection process. It includes several images used for the analysis, along with tabulated PSNR values, obtained during compression using various algorithms. Section 3 presents the results of the face detection process using the Haar Cascade Classifier for frontal and non-frontal compressed images. The results are represented in tabular form. The tables contain F - the numerical values for accuracy (F-measure/F-score) and Det.F - the total count of detected faces (true and false), as well as the number of correctly detected faces, written in parenthesis (when there are additional false detections). Finally, Section 4 provides the conclusion and key insights, along with potential directions for future research.



2. SYSTEM MODEL

This paper analyzes the impact of the JPEG, JPEG2000, and SPIHT algorithms on face detection using the Haar Cascade Classifier. The images were extracted from the FDD database [12], along with the corresponding GroundTruth, which defines the exact position of every facial component [13]. All three compression techniques were applied to each image using the following bits-per-pixel (BPP) values: 0.1, 0.2, 0.4, 0.5, 1, and 1.5. The selected images met the criteria for the number of faces represented, specifically 1, 3, 5, and 10, along with the perspective criteria of frontal and non-frontal. Each image represented a different number of faces captured from various perspectives.

The images extracted from the FDD database were converted to BMP (bitmap) format using Python code, allowing the image compression to be performed using VcDemo software [14]. Face detection was performed utilizing OpenCV's Haar Cascade Classifier, serving as an essential element of the Python code.

The images used for the analysis, the obtained images, and the codes were all stored in a repository [15] created by the authors of this paper.

The following objective measures were used:

F - the evenly balanced F-measure that is calculated using the following equation [16] [17]:

$$F = \frac{2PR}{P + R}$$

Equation 1. Calculation of F-measure

Where P represents Precision, and R represents Recall or Sensitivity [16].

In an ideal situation, F is equal to 1, but, generally, F lies within the limits of $0 \leq F \leq 1$ [1].

The Precision is calculated using the following equation [17]:

$$P = \frac{TP}{TP + FP}$$

Equation 2. Calculation of Precision

The Recall is calculated using the following equation [17]:

$$R = \frac{TP}{TP + FN}$$

Equation 3. Calculation of Recall

Where: TP represents True Positive - the pixels correctly segmented as foreground; FP represents False Positive - the pixels falsely segmented as foreground; TN represents True Negative - the pixels correctly detected as background, and FN represents False Negative - the pixels falsely detected as background [1].

Det.F - the total count of detected faces (true and false), with the number of correctly detected faces written in parenthesis (when there are additional false detections).

Peak signal-to-noise ratio (PSNR) values for the three compression algorithms used in multi-face images are represented in the following tables. Table 1 presents the PSNR values obtained from compressing images containing either one face or three faces; Table 2 presents the PSNR values obtained from compressing images containing either five faces or ten faces.

It's readily observed how bits-per-pixel (BPP) values influence image compression. Higher value of BPP results in a better image quality, especially with JPEG compression. JPEG2000 and SPIHT achieve similar results, but JPEG2000 achieves slightly better PSNR values at lower BPP and SPIHT at higher BPP values.

Table 1. PSNR values for the three compression algorithms at different BPP values for frontal and non-frontal images containing one face and three faces

Perspective	Compression Algorithm	One Face						Three Faces					
								BPP value					
		0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5
Frontal	JPEG	24.6	40.7	44.6	48	55.3	56	23.6	33.5	38.3	39.3	42.4	52
	JPEG2000	42.3	45.1	48.4	49.5	51.1	51.1	35.1	38.4	41.3	42.1	45.5	48.5
	SPIHT	42.7	45.5	48.9	50.1	54.2	57.3	35	38.4	41.3	42.3	45.8	49
Non-frontal	JPEG	23.5	31.2	35	36.2	39.6	50.5	24.9	35.7	39.8	40	54.3	56.8
	JPEG2000	32.5	35.2	38.3	39.7	45.4	49.4	36.6	39.1	42.4	43.7	48.8	51.1
	SPIHT	32.1	35	38.3	39.7	45.6	50	36.9	39.5	42.7	44	49.4	53.7



Table 2. PSNR values for the three compression algorithms at different BPP values for frontal and non-frontal images containing five faces and ten faces

Perspective	Compression Algorithm	Five Faces						Ten Faces					
		BPP value											
		0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5
Frontal	JPEG	21.9	26.9	33.3	34.8	39.7	42.4	21.2	25.1	29.3	31	34.4	44.1
	JPEG2000	28.5	32.2	36.4	37.9	43.2	46.6	25.8	28.6	32.3	33.8	39.3	43.4
	SPIHT	28.1	32	36.3	37.7	43.1	46.7	25.4	28	31.9	33.4	39.1	43.3
Non-frontal	JPEG	22.9	30.1	37	38.6	42.9	45.7	21.5	26.1	31.9	33.5	38.4	42.1
	JPEG2000	32	36	40.4	41.7	45.8	48.5	27.3	31	35.6	37.2	43.1	46.9
	SPIHT	31.6	35.2	40	41.4	45.8	48.7	26.3	30.1	35.1	36.9	42.9	47.1

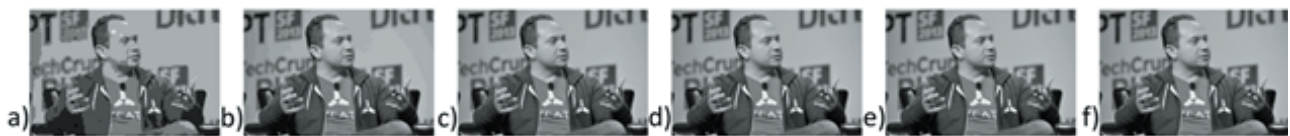


Figure 1. Non-frontal images representing one face with JPEG compression at the BPP value of:
a) 0.1, b) 0.2, c) 0.4, d) 0.5, e) 1, f) 1.5

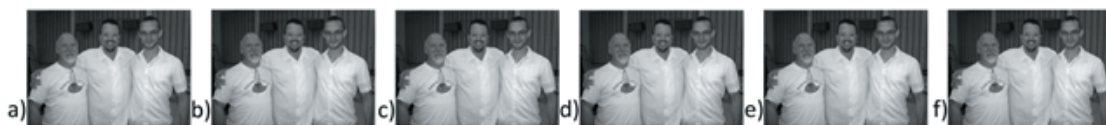


Figure 2. Frontal images representing three faces with JPEG2000 compression at the BPP value of:
a) 0.1, b) 0.2, c) 0.4, d) 0.5, e) 1, f) 1.5

Several images resulting from the compression processes are shown below.

Figure 1 shows compressed, non-frontal images for different values of BPP, representing one face when using JPEG compression. Figure 2 shows compressed, frontal images for different values of BPP, representing three faces when using JPEG2000 compression. The rest of the compressed images used in this work are available in the dataset repository [15].

Figure 1 and Figure 2 show that image quality remains acceptable across various compression methods, but lower BPP values and certain compression algorithms significantly affect image quality. The most noticeable loss in quality occurs at low BPP values when using the JPEG compression. This is confirmed by the data in Table 1.

In 2001, Paul Viola and Michael Jones introduced the Viola-Jones object detection framework [9]. This framework is the first to achieve competitive real-time

detection performance. Due to its accuracy and efficiency, it has been used in face detection ever since. It is important to note that the primary goal of this framework is face detection, not recognition. This detection phase serves as an elemental processing step that comes before any subsequent face recognition processes [9].

The Viola-Jones face detection algorithm represents the core foundation of the Haar Cascade Classifier. Training this classifier requires a sufficient number of both positive and negative images. It uses Haar-like features. A specific object in an image can be detected using Haar-like features. The eyes, nose, and mouth, as the integral parts of a human face, represent features that can be used for classification. By comparing calculated and trained feature values (the reference values), the Haar Cascade Classifier gains the ability to detect human faces [18] [19].

The following section presents the results obtained during the face detection process.



3. RESULTS

This section presents the results of the face detection processes. These results were gathered using the mathematical models and software tools outlined in Section 2. Table 3, Table 4, Table 5, and Table 6 display the F and Det.F values for frontal and non-frontal images, based on the number of represented faces. These values were obtained by running a Python code with the Haar Cascade Classifier as its fundamental part, changing digital image inputs based on the number of faces, perspective, and applied compression algorithms.

As shown in Table 3, it is evident that for images containing a single frontal face, the F-score consistently equals 1, indicating perfect face detection, regardless of the BPP value or the applied compression algorithm; the results are almost identical for a single non-frontal face,

with only one exception: in the image compressed using the JPEG algorithm at the BPP value of 0.1, where no faces were detected.

Table 4 presents data indicating perfect face detection in compressed images representing three frontal faces, using JPEG2000 compression standard, regardless of the BPP value. The situation is slightly different for the remaining two compression algorithms, with the F-score falling below 1 for the BPP value of 1.5 applying JPEG and for the BPP values of 1 and 1.5 applying the SPIHT algorithm, showing subtle inverse proportionality characteristics. In the case of non-frontal images representing three faces, the results suggest that, for most of the BPP value cases, face detection using the Haar Cascade Classifier achieves the same results - the F-score of 0.67, with two true detections and a false one. There are a few exceptions, with the worst case for the BPP

Table 3. Accuracy and number of detected faces using the Haar Cascade Classifier for face detection in frontal and non-frontal images representing one face, compressed by various methods at different BPP values

Frontal images												Non-frontal images											
JPEG												JPEG											
0.1		0.2		0.4		0.5		1		1.5		0.1		0.2		0.4		0.5		1		1.5	
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
JPEG2000												JPEG2000											
0.1		0.2		0.4		0.5		1		1.5		0.1		0.2		0.4		0.5		1		1.5	
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SPIHT												SPIHT											
0.1		0.2		0.4		0.5		1		1.5		0.1		0.2		0.4		0.5		1		1.5	
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 4. Accuracy and number of detected faces using the Haar Cascade Classifier for face detection in frontal and non-frontal images representing three faces, compressed by various methods at different BPP values

Frontal images												Non-frontal images											
JPEG												JPEG											
0.1		0.2		0.4		0.5		1		1.5		0.1		0.2		0.4		0.5		1		1.5	
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
1	3	1	3	1	3	1	3	1	3	0.86	4(3)	0.57	4(2)	0.8	2(2)	0.67	3(2)	0.67	3(2)	0.67	3(2)	0.67	3(2)
JPEG2000												JPEG2000											
0.1		0.2		0.4		0.5		1		1.5		0.1		0.2		0.4		0.5		1		1.5	
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
1	3	1	3	1	3	1	3	1	3	1	3	0.67	3(2)	0.67	3(2)	0.67	3(2)	0.67	3(2)	0.67	3(2)	0.67	3(2)
SPIHT												SPIHT											
0.1		0.2		0.4		0.5		1		1.5		0.1		0.2		0.4		0.5		1		1.5	
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
1	3	1	3	1	3	1	3	0.86	4(3)	0.86	4(3)	0.33	3(1)	0.67	3(2)	0.67	3(2)	0.67	3(2)	0.67	3(2)	0.67	3(2)



value of 0.1 with SPIHT, the second worst for with BPP value of 0.1 with JPEG, and the best for the BPP value of 0.2 with JPEG.

Referring to images representing five frontal faces, Table 5 shows that the JPEG compression algorithm achieves both the best and worst results, with an F-score of 1 for the BPP value of 0.1 and an F-score of 0.83 for the BPP value of 0.2 (2 additional false detections). For the remaining cases of BPP values, JPEG achieves an identical F-score to all the BPP cases with both JPEG2000 and SPIHT algorithms, where the F-score has the same value of 0.91 (1 additional false detection). Therefore, it is observed that JPEG2000 and SPIHT achieve identical, consistent F-scores regardless of the BPP value, while there are slight deviations in accuracy when using JPEG. The results for non-frontal images representing five faces showed an F-score of 0.75 (3 correct detections) for al-

most all the BPP values and using all three compression methods, except with SPIHT at the BPP value of 0.1, which represents the worst scenario of Haar Cascade-based face detection in non-frontal images representing five faces, achieving the F-score of 0.67 (4 detections, 3 correct).

Observing data represented in Table 6, the first situation of unclear conclusion and absence of BPP to F-score relation occurs. Regarding compressed images representing ten frontal faces, from the results obtained, it is seemingly impossible to find a causal-consequential relationship between BPP values and the F-score. All three algorithms provide a case of perfect face detection, JPEG for the BPP value of 0.1, and both JPEG2000 and SPIHT for the BPP value of 0.4, and a case of worst F-score of 0.87 (3 additional false detections) for the BPP value of 0.2. The rest of the F-scores, for all the remaining BPP

Table 5. Accuracy and number of detected faces using the Haar cascade classifier for face detection in frontal and non-frontal images representing five faces, compressed by various methods at different BPP values

Frontal images												Non-frontal images											
JPEG												JPEG											
0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
1	5	0.83	7(5)	0.91	6(5)	0.91	6(5)	0.91	6(5)	0.91	6(5)	0.75	3(3)	0.75	3(3)	0.75	3(3)	0.75	3(3)	0.75	3(3)	0.75	3(3)
JPEG2000												JPEG2000											
0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
0.91	6(5)	0.91	6(5)	0.91	6(5)	0.91	6(5)	0.91	6(5)	0.91	6(5)	0.75	3(3)	0.75	3(3)	0.75	3(3)	0.75	3(3)	0.75	3(3)	0.75	3(3)
SPIHT												SPIHT											
0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
0.91	6(5)	0.91	6(5)	0.91	6(5)	0.91	6(5)	0.91	6(5)	0.91	6(5)	0.67	4(3)	0.75	3(3)	0.75	3(3)	0.75	3(3)	0.75	3(3)	0.75	3(3)

Table 6. Accuracy and number of detected faces using the Haar Cascade Classifier for face detection in frontal and non-frontal images representing three faces, compressed by various methods at different BPP values

Frontal images												Non-frontal images											
JPEG												JPEG											
0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
1	10	0.87	13(10)	0.95	11(10)	0.91	12(10)	0.95	11(10)	0.91	12(10)	0.18	1(1)	0.29	4(2)	0.4	5(3)	0.15	3(1)	0.29	4(2)	0.27	5(2)
JPEG2000												JPEG2000											
0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
0.91	12(10)	0.87	13(10)	1	10	0.95	11(10)	0.95	11(10)	0.91	12(10)	0.15	3(1)	0.29	4(2)	0.37	6(3)	0.4	5(3)	0.37	6(3)	0.29	4(2)
SPIHT												SPIHT											
0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5	0.1	0.2	0.4	0.5	1	1.5
F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F	F	Det.F
0.95	11(10)	0.87	13(10)	1	10	0.95	11(10)	0.95	11(10)	0.91	12(10)	0.17	2(1)	0.29	4(2)	0.4	5(3)	0.4	5(3)	0.37	6(3)	0.29	4(2)



value cases in all the applied compression algorithms, switch between 0.91 (2 additional false detections) and 0.95 (1 additional false detection) without discernible causality. A large number of non-frontal faces in a compressed image leads to poor face detection results when using the Haar Cascade Classifier, with the highest F-score being as low as 0.4.

Examining the obtained results, it is noticeable that there are several characteristic cases of Haar Cascade-based face detection in images representing faces from the frontal perspective, as well as in images representing faces from the non-frontal perspective. The following figures show the above-mentioned characteristic cases. Figure 3 displays JPEG images with faces captured from the frontal perspective and compressed at the BPP value of 0.1, showing perfect face detection regardless of the number of faces represented. Figure 4 displays images with ten faces captured from both frontal and non-frontal perspectives and compressed using JPEG, JPEG2000, and SPIHT algorithms at various BPP values, showing the worst face detection results.

The remaining face detection images generated in this study are available in the dataset repository [15].

By gathering all the results from the processes detailed in Section 2 and Section 3, the following conclusions were drawn.

4. CONCLUSION

This paper presents an analysis of face detection using the Haar Cascade Classifier in frontal and non-frontal images representing different numbers of faces and compressed using JPEG, JPEG2000, and SPIHT algorithms at different bits-per-pixel values.

The obtained results indicate that the Haar Cascade Classifier does a significantly better job performing face detection in images showing frontal faces, regardless of the applied compression algorithm. Given the collected data, the described face detection method achieves the perfect F-score of 1 in frontal JPEG images for the BPP value of 0.1. There aren't cases of undetected faces in frontal face images, regardless of the compression algorithm. Single-face images have the best chance of getting the perfect F-score; as the number of faces in an image increases, the F-score remains high but becomes more unpredictable, with no distinct connection to the BPP value.

The data collected for the non-frontal face images confirms that the Haar Cascade Classifier is less effective when face detection is needed in non-frontal images, with better results being achieved in images representing smaller numbers of faces and the F-score dropping in an inverse proportionality manner to the rising number of represented faces. Additionally, there's an observed relationship between the angle of rotation of the face and the face detection accuracy, with a smaller angle of rotation increasing the chance of a correct face detection.



Figure 3. Frontal JPEG images compressed at the BPP value of 0.1, showing perfect face detection regardless of the number of faces represented



Figure 4. Frontal a) JPEG, b) JPEG2000, and c) SPIHT images compressed at the BPP value of 0.2, and non-frontal d) JPEG, e) JPEG2000, and f) SPIHT images compressed at the highest BPP value of 1.5, representing ten faces, showing the worst results of face detection



The findings contribute to the further development of image compression algorithms to make them more suitable for use in face detection systems and, also, to the necessity of additional improvements in the Haar Cascade Classifier, especially in cases with a larger number of faces and greater angles of perspective. Finding an approach for solving these problems indicates a possible direction for future research.

REFERENCE

- [1] V. Maksimović, B. Jakšić, M. Petrović, P. Spalević and M. Milošević, "Analysis of Edge Detection on Compressed Images with Different Complexities," *Acta Polytechnica Hungarica*, vol. 2, no. 2, pp. 77-81, 2020.
- [2] A. Subramanya, "Image compression technique," *IEEE potentials*, vol. 20, no. 1, pp. 19-23, 2001.
- [3] A. A. M. Kumar, "An Introduction to Image Compression," *International Journal of Computer Science and Information Technology Research*, vol. 2, no. 2, pp. 77-81, 2014.
- [4] G. Huson, A. Léger, B. Niss and I. Sebestyén, "JPEG at 25: Still going strong," *IEEE MultiMedia*, vol. 24, no. 2, pp. 96-103, 2017.
- [5] G. A. Papakostas, E. G. Karakasis and D. E. Koulouriotis, "On accelerating the computation of 2-D discrete cosine transform in image processing," in *ICSES 2008 International Conference on Signals and Electronic Systems Proceedings*, Krakow, Poland, 2008.
- [6] D. Taubman, M. W. Marcellin and M. Rabbani, "JPEG2000: Image compression fundamentals, standards and practice," *Journal of Electronic Imaging*, vol. 11, no. 2, pp. 286-287, 2002.
- [7] C. Christopoulos, A. Skodras and T. Ebrahimi, "The JPEG2000 still image coding system: an overview," *IEEE transactions on consumer electronics*, vol. 46, no. 4, pp. 1103-1127, 2000.
- [8] P. D. Saraf, D. Sisodia, A. Sinhal and S. Sahu, "Design and implementation of novel SPIHT algorithm for image compression," in *Proceedings of 2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Solan, India, 2012.
- [9] K. Dang and S. Sharma, "Review and comparison of face detection algorithms," in *Proceedings of the 7th International Conference Confluence on Cloud Computing, Data Science & Engineering*, Noida, India, 2017.
- [10] P. Elmer, A. Lupp, S. Sprenger, R. Thaler and A. Uhl, "Exploring compression impact on face detection using haar-like features," in *Image Analysis: 19th Scandinavian Conference, SCIA 2015, Copenhagen, Denmark, June 15-17, 2015. Proceedings*, Copenhagen, Denmark, 2015.
- [11] N. Prasad, B. Rajpal, K. K. Mangalore, R. Shastri and N. Pradeep, "Frontal and non-frontal face detection using deep neural networks (DNN)," *International Journal of Research in Industrial Engineering*, vol. 10, no. 1, pp. 9-21, 2021.
- [12] F. Elmenshawii, "Kaggle," 10 6 2023. [Online]. Available: <https://www.kaggle.com/datasets/fare-selmenshawii/face-detection-dataset>. [Accessed 10 1 2025].
- [13] R. Queiroz, M. Cohen, J. L. Moreira, A. Braun, J. C. J. Júnior and S. R. Musse, "Generating facial ground truth with synthetic faces," in *2010 23rd SIBGRAPI - Conference on Graphics, Patterns and Images*, Gramado, Brazil, 2010.
- [14] Delft University of Technology (TU-Delft), "TUDelft," ICT Group, TU-Delft, 13 9 2020. [Online]. Available: <https://homepage.tudelft.nl/c7c8y/VcDemo.html>. [Accessed 9 1 2025].
- [15] I. Šarkočević, "Google Drive: Dataset Repository - Haar Cascade-based Face Detection," 4 4 2025. [Online]. Available: <https://drive.google.com/drive/folders/1Yt-x6uNf8INYEmJsuMlMyX6dRLRH0XKC?usp=sharing>. [Accessed 4 4 2025].
- [16] Y. Sasaki, "The truth of the F-measure," *Teach tutor mater*, vol. 1, no. 5, pp. 1-5, 2007.
- [17] M. Sokolova, N. Japkowicz and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence*, Hobart, Australia, December 4-8, 2006, Proceedings, Hobart, Australia, 2006.
- [18] S. K. Mondal, I. Mukhopadhyay and S. Dutta, "Review and comparison of face detection techniques," in *Proceedings of International Ethical Hacking Conference 2019: eHaCON 2019*, Kolkata, India, Kolkata, India, 2020.
- [19] C. H. Choi, J. Kim, J. Hyun, Y. Kim and B. Moon, "Face detection using haar cascade classifiers based on vertical component calibration," *Human-centric Computing and Information Sciences*, vol. 12, no. 11, pp. 1-17, 2022.



DATA SCIENCE AND APPLICATIONS SESSION





FORECASTING BASE METALS PRICES: A COMPARISON OF VARIOUS BAYESIAN-BASED METHODS

Krzysztof Drachal*,
[0000-0001-5297-6695]

Joanna Jędrzejewska
[0009-0006-4471-9078]

Faculty of Economic Sciences,
University of Warsaw,
Warsaw, Poland

Abstract:

This paper addresses the topic of forecasting base metal prices index using advanced Bayesian methods, emphasising Bayesian dynamic mixture models. Original schemes were expanded by certain modifications. A broad set of macroeconomic indicators, such as interest rates, industrial production, economic activity, market stress indices, others commodities prices, exchange rates and information from stock markets, etc. were taken as potential predictors. Models were recursively estimated, taking under consideration possible discrepancy between released and revised data, carefully simulating real-time forecasting conditions. Dynamic Model Averaging was found to provide the highest accuracy of predictions compared to competing models. The forecasts were significantly more accurate than the ARIMA method or the no-change method. Among the dynamic mixture variants, model selection appeared to offer the best performance. The Clark-West test for nested models confirmed that forecast combination schemes lead to significant forecast accuracy improvements. Sector companies' stock prices and particular exchange rates were found to be the important base metals price predictors.

Keywords:

Bayesian Dynamic Mixture Models, Dynamic Model Averaging, Forecasting Accuracy, Model Averaging, Model Selection.

INTRODUCTION

Forecasting metal prices has become an important area of focus in economic research. Over time, there has been a shift from traditional econometric methods to more advanced techniques that promise greater prediction accuracy. Precise forecasting of metal prices is vital for various stakeholders, such as policymakers, investors, and industries that rely on raw materials. Metal prices are affected by numerous factors, including supply and demand fluctuations, geopolitical developments, and financial market conditions, which makes predicting them a complex task. Recently, the adoption of advanced methods, especially machine learning models, has gained traction in the field of metal price forecasting [1-5]. The aim of this research is to provide an insight into this topic from Bayesian dynamic mixture models (BDMM) applied to forecasting the monthly World Bank index of metal prices [6]. Contrary to traditional models often failing to account for non-linear relationships and structural breaks caused by unpredictable events, such as the COVID-19 pandemic or geopolitical tensions, BDMM offers a promising technique [7-8].

Correspondence:

Krzysztof Drachal

e-mail:

kdrachal@wne.uw.edu.pl





2. DATA

Monthly data between 09/2002 and 10/2024 were analysed. World Bank Base Metals Price Index (2010 = 100, USD), including aluminium, copper, lead, nickel, tin and zinc was taken as the dependent variable. Similarly, as in previous studies [1-5], and, for example, Buncic and Moretto [9] the following variables were taken as independent ones: Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCSL); U.S. 3-month risk-free rate (monthly averages, 3-month treasury bill: secondary market rate, TB3MS), a proxy of short-term interest rate; U.S. 10-year government bond yields (monthly averages, IRLTLT01USM156N), a proxy of long term interest rate; U.S industrial production (INDPRO); Kilian Index of Global Real Economic Activity (IGREA); term spread (10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity for U.S.; monthly averages, T10Y2YM); St. Louis Fed Financial Stress Index (monthly averages, STLFSI4); implied volatility VIX index (monthly averages); The Caldara and Iacoviello GPR index; S&P 500 index ($\wedge$ SPX); Dow Jones Industrial index ($\wedge$ DJI); Shanghai Composite Index ($\wedge$ SHC); MSCI EM index for emerging markets; monthly average prices of gold (USD per troy ounce); monthly average WTI oil price (USD per barrel). Additionally, exchange rates (to USD) of the largest producers of aluminium, lead, nickel and zinc (both primary metal and ore) were taken, i.e., China (USDCNY), Russia (USDRUB), India (USDINR), Indonesia (CCUSSP02IDM650N and since 01/2024 US-DIDR), South Korea (USDKRW), Australia (AUDUSD), Guinea (monthly averages, GNF) and Philippines (USD-PPH). Also, largest metal companies share prices were taken, i.e., Rio Tinto Plc (RIO.UK, in GBX), Alcoa Corp (AA.US, in USD), Hindustan Zinc Limited (HINDZINC.BO, in INR), Teck Resources Limited (TECK, in USD), BHP Group Limited (BHP, in USD), and Sherritt International Corporation (S.TO, in CAD) [6, 10-22].

If not otherwise stated, the last observations in a period were taken. All variables, except interest rates, Kilian index, term spread, financial stress index, VIX and GPR were included in log-differences. For computational reasons, and stationarity issues, the variables were later standardized based on a mean and standard deviation of the first 100 observations (Table 1 and Table 2). This value was kept further as the in-sample. However, obtained forecasts were transformed back to the level values, and then evaluated. To mimic real-market data availability, all independent variables were lagged one period back, and the Kilian index was lagged two periods back. Initially, the data as released in the past were taken, i.e., not in a revised form [23].

3. METHODS

All models were estimated recursively, i.e., each forecast at time t was done over all the data available up to time $t-1$. The following models were estimated: Bayesian dynamic mixture (BDMM) with state space (SS) and normal regression components (NR). The original schemes [7-8] were additionally improved by model averaging (A) and model selection schemes (H) [24]. They were denoted by BDMM-SS, BDMM-SS-A, BDMM-SS-H, BDMM-NR and BDMM-NR-H respectively. Additionally, Dynamic Model Averaging (DMA) and Dynamic Model Selection (DMS) were estimated (with the standard forgetting factors equal to 0.99). Bayesian Model Averaging (BMA) and Bayesian Model Selection (BMS) were also estimated [25]. These models were also estimated in versions with component models being just single variable ones (i.e., simple linear regression of one independent variable and a constant). They were denoted by DMA-1VAR, DMS-1VAR, BMA-1VAR and BMS-1VAR, respectively [26].

LASSO, RIDGE and Elastic Net (EL-NET) regressions were estimated, both conventionally and in their Bayesian versions, i.e., B-LASSO and B-RIDGE. Following the recursive approach, the penalty parameter was chosen by the t -fold cross-validation based on the Mean Square Error (MSE) metric. Mixing parameters {0.1, 0.2, ..., 0.9} were employed. Additionally, the least-angle regression (LARS) was estimated [27-29].

Time-varying parameters regressions, both with forgetting (TVP-FOR) and without (TVP) were estimated [26]. ARIMA was estimated with automatic lag selection [30]. The no-change (NAÏVE) method, a historical average over all past observations (HA) and over rolling window of last 100 observations (HA-ROLL) were also estimated.

In addition, when suitable, variance updating was done as by Raftery et al. [25], and additionally the version with the exponentially weighted moving average was estimated with parameter $\kappa = 0.97$ as recommended, for example, in [31]. Such version of models was denoted by adding "-K", i.e., BDMM-SS-K, BDMM-SS-A-K, BDMM-SS-H-K, DMA-K, DMS-K, DMA-1VAR-K, DMS-1VAR-K, BMA-K, BMS-K, BMA-1VAR-K, BMS-1VAR-K, TVP-K and TVP-FOR-K. Secondly, BDMM-NR-1MOD denotes the model BDMM-NR where the only component is the model with all considered independent variables. In the case of BDMM-SS and BMA such a reduction results in the TVP model, and for DMA – in TVP-FOR model [26].



Table 1. Descriptive statistics

	Min	Max	Mean	Median	Standard Deviation	Coefficient of Variation	Skewness
p_metals	-4.2048	2.1591	-0.1208	-0.1038	0.7820	-6.4713	-0.8220
cpi	-4.8579	2.8450	0.0230	-0.0025	0.8613	37.4467	-0.6132
r_short	-1.1416	1.9272	-0.2644	-0.6292	1.0223	-3.8662	0.9867
r_long	-5.3409	1.7318	-1.5594	-1.7494	1.7734	-1.1372	-0.0638
ind_prod	-8.0029	1.9768	0.0960	0.1892	0.7306	7.6086	-6.7462
ec_act	-4.2736	2.0267	-1.1602	-1.2689	1.2785	-1.1019	0.3432
term_spread	-2.3669	1.3546	-0.3392	-0.2884	0.9734	-2.8695	-0.0524
fin_stress	-0.8724	4.8788	-0.2141	-0.4022	0.6875	-3.2116	3.7128
VIX	-1.0883	4.0797	-0.1901	-0.3819	0.8051	-4.2355	2.3035
GPR	-0.7115	5.7244	0.1868	0.0023	0.8288	4.4357	2.8961
SP500	-4.1981	2.5353	0.0614	0.1757	0.9446	15.3745	-0.7754
DJ_Ind	-3.6352	2.9383	0.0453	0.1269	0.9665	21.3151	-0.5550
SSE	-3.1314	2.5712	-0.0318	-0.0151	0.7847	-24.7042	-0.4833
MSCI	-4.6947	1.9523	-0.1250	-0.1160	0.8331	-6.6672	-0.8312
p_gold	-3.3274	2.3270	-0.1542	-0.2054	0.8840	-5.7330	-0.0410
p_oil	-6.4683	5.7477	-0.0826	0.0619	1.1254	-13.6287	-0.9461
fx_CNY	-3.1334	4.6269	0.1768	0.1050	1.2344	6.9823	0.5422
fx_RUB	-5.4840	7.6815	0.1817	0.0160	1.7863	9.8299	0.8612
fx_INR	-3.6819	4.1998	0.1285	0.0430	1.0303	8.0189	0.2759
fx_IDR	-3.3691	5.1459	0.0662	0.0581	0.8740	13.2031	0.9473
fx_KRW	-3.9837	3.3854	0.0337	-0.0188	0.8094	24.0421	0.1544
fx_AUD	-4.4667	2.3168	-0.1323	-0.0929	0.8666	-6.5500	-0.5371
fx_GNF	-6.3337	3.2157	-0.1344	-0.2190	0.6782	-5.0449	-2.0818
fx_PHP	-2.2100	2.6953	0.1191	0.0638	0.9346	7.8471	0.2739
Rio_Tinto	-5.0927	2.1439	-0.0625	-0.0789	0.8114	-12.9911	-1.0748
Alcoa	-6.3662	3.4061	0.0121	0.0586	1.1152	92.1101	-1.1646
Hindustan_Zinc	-2.4213	3.6508	-0.1208	-0.1596	0.7362	-6.0935	0.8719
Teck	-5.5518	3.0392	-0.0918	-0.1007	0.8403	-9.1537	-0.9763
BHP	-3.3694	2.5499	-0.1436	-0.1179	0.9520	-6.6297	-0.2834
Sheritt	-5.4259	4.0323	-0.1377	-0.1582	1.1137	-8.0889	-0.4420



Table 2. Stationarity tests. (For all tests 0.0100 indicates value less than 0.01. For Kwiatkowski-Phillips-Schmidt-Shin test 0.1000 indicates value greater than 0.1.)

	augmented Dickey-Fuller test statistic	augmented Dickey-Fuller test p-value	Phillips- Perron test statistic	Phillips- Perron test p-value	Kwiatkowski- Phillips- Schmidt-Shin test statistic	Kwiatkowski- Phillips- Schmidt-Shin test p-value	Skewness
p_metals	-6.2814	0.0100	-173.3586	0.0100	0.1324	0.1000	-0.8220
Cpi	-4.8201	0.0100	-133.1739	0.0100	0.3016	0.1000	-0.6132
r_short	-2.6648	0.2959	-2.8191	0.9420	0.5390	0.0329	0.9867
r_long	-1.0857	0.9234	-6.2037	0.7623	1.8374	0.0100	-0.0638
ind_prod	-7.1646	0.0100	-228.9841	0.0100	0.2105	0.1000	-6.7462
ec_act	-2.8170	0.2318	-22.9402	0.0369	1.7232	0.0100	0.3432
term_spread	-2.2485	0.4713	-6.6974	0.7346	1.4911	0.0100	-0.0524
fin_stress	-3.3587	0.0618	-31.1382	0.0100	0.3217	0.1000	3.7128
VIX	-3.4809	0.0450	-41.8297	0.0100	0.1830	0.1000	2.3035
GPR	-7.0251	0.0100	-64.7243	0.0100	0.7558	0.0100	2.8961
SP500	-5.7640	0.0100	-257.3397	0.0100	0.1015	0.1000	-0.7754
DJ_Ind	-5.6475	0.0100	-253.0829	0.0100	0.0644	0.1000	-0.5550
SSE	-4.7156	0.0100	-271.5288	0.0100	0.0420	0.1000	-0.4833
MSCI	-6.1156	0.0100	-245.3847	0.0100	0.2399	0.1000	-0.8312
p_gold	-5.6370	0.0100	-221.1622	0.0100	0.2435	0.1000	-0.0410
p_oil	-6.5283	0.0100	-168.1905	0.0100	0.0775	0.1000	-0.9461
fx_CNY	-6.5491	0.0100	-254.2642	0.0100	0.3558	0.0962	0.5422
fx_RUB	-6.6391	0.0100	-195.8817	0.0100	0.1502	0.1000	0.8612
fx_INR	-5.5961	0.0100	-233.7721	0.0100	0.1668	0.1000	0.2759
fx_IDR	-6.6932	0.0100	-238.7224	0.0100	0.0577	0.1000	0.9473
fx_KRW	-6.3087	0.0100	-270.5470	0.0100	0.1184	0.1000	0.1544
fx_AUD	-6.5947	0.0100	-258.6252	0.0100	0.2818	0.1000	-0.5371
fx_GNF	-5.8932	0.0100	-137.3176	0.0100	0.3462	0.1000	-2.0818
fx_PHP	-6.1184	0.0100	-248.4363	0.0100	0.2890	0.1000	0.2739
Rio_Tinto	-6.7902	0.0100	-249.9814	0.0100	0.0558	0.1000	-1.0748
Alcoa	-5.7666	0.0100	-279.7361	0.0100	0.0475	0.1000	-1.1646
Hindustan_Zinc	-5.8495	0.0100	-230.7092	0.0100	0.3541	0.0970	0.8719
Teck	-7.3485	0.0100	-236.2602	0.0100	0.0777	0.1000	-0.9763
BHP	-6.3720	0.0100	-266.9019	0.0100	0.2585	0.1000	-0.2834
Sherritt	-6.4427	0.0100	-240.5759	0.0100	0.1512	0.1000	-0.4420



Herein, $K=29$ independent variables are considered. In case of BDMM schemes and DMA, DMS, BMA and BMS, all possible multilinear regression models (i.e., 2^K , because the constant-only model is also included) are originally considered as component models. This leads to serious computational obstacles. Therefore, DMA-1VAR, DMS-1VAR, BMA-1VAR and BMS-1VAR models were estimated, as they are based on $K+1$ models only (the model with constant only is included). The second approach to reduce the number of component models is as follows. The set of independent variables is split into interest rates, economic activity, market stress, stock markets, other commodities prices, exchange rates, and metal sector stock price factors, i.e.,

$x_1 = [\text{cpi}, r_{\text{short}}, r_{\text{long}}];$

$x_2 = [\text{ind\_prod}, \text{ec\_act}];$

$x_3 = [\text{term\_spread}, \text{fin\_stress}, \text{VIX}, \text{GPR}];$

$x_4 = [\text{SP500}, \text{DJ\_Ind}, \text{SSE}, \text{MSCI}];$

$x_5 = [\text{p\_gold}, \text{p\_oil}];$

$x_6 = [\text{fx\_CNY}, \text{fx\_RUB}, \text{fx\_INR}, \text{fx\_IDR}, \text{fx\_KRW}, \text{fx\_AUD}, \text{fx\_GNF}, \text{fx\_PHP}];$ and

$x_7 = [\text{Rio\_Tinto}, \text{Alcoa}, \text{Hindustan\_Zinc}, \text{Teck}, \text{BHP}, \text{Sherritt}].$

Then, all possible multilinear regression models are constructed for each set of independent variables, i.e., for $x_1, x_2, \dots, x_7$. Finally, the model with all independent variables is added. As a result, instead of 2^{29} , just $1 + 2^3 + 2^2 + 2^4 + 2^4 + 2^2 + 2^8 + 2^6 - 7 + 1 = 363$ component models must be considered, which corresponds to less than 9 independent variables for the original scheme, and which is computationally feasible. Simultaneously, such a split and combinations are economically reasonable, as they keep forecast averaging idea of the modelling schemes, and emphasise different economic groups of factors possibly influencing metals prices.

4. RESULTS

Table 3 presents forecast accuracy metrics of the estimated models. Root Mean Square Error (RMSE), normalized RMSE (N-RMSE), Mean Absolute Error (MAE) and Mean Absolute Scaled Error (MASE) were computed [32]. According to all metrics, DMA-K is the most accurate method, followed by DMA. Despite the poor performance of BDMM models, several of these schemes outperform NAÏVE or ARIMA. Out of these schemes, BDMM-NR-H performs the best. If revised data is taken, outcomes are quite comparable [23]. Indeed, for the robustness of results, initially, the models

were estimated with released data, mimicking real-time forecasting. However, versions with revised data (as of 12/2024) were also estimated [10-11, 23].

According to the Diebold-Mariano test [33] with 10% significance level and squared error loss function, forecasts from DMA-K are significantly more accurate than those from NAÏVE and ARIMA, TVP, TVP-FOR, TVP-K and TVP-FOR-K, as well as, many other models, but not from BDMM-SS-A-K or BDMM-NR-H. On the other hand, BDMM-NR-H forecasts more accurately than NAÏVE, but not than ARIMA. It forecasts more accurately than BDMM-NR-1MOD, and more accurately than many other models, but not as much as DMA-K. If revised data is taken, outcomes are similar, but BDMM-NR-H is found to be additionally more accurate than ARIMA. (Due to the limited space detailed outcomes are not presented herein.) Moreover, the Model Confidence Set test [34], at a 5% significance level, eliminated BDMM-SS, BDMM-SS-K, TVP, HA and HA-ROLL models.

When forecasts from models with released data were compared with those based on revised data with the Diebold-Mariano test with a 5% significance level, both with squared error and absolute scaled error loss functions, different accuracy can be assumed for BDMM-SS, RIDGE and TVP. For EL-NET, B-LASSO, B-RIDGE and TVP-K different accuracy can be assumed only when absolute scaled error loss functions are applied.

The Giacomini-Rossi fluctuation test over approximately 2.75-year periods, at a 5% significance level, does not indicate that BDMM-NR-H forecasting performance would perform worse than DMA-K [35].

The Clark-West test for nested models was performed [36]. Only in two cases, assuming 5% significance level, the larger model cannot be said to generate smaller errors than the restricted (simple) model. This provides some evidence that model combination schemes provide significant gains to increase forecast accuracy. Details are in Table 4.

Finally, in case of variable selection, DMA-K ascribed the highest weights to share prices, whereas BDMM-NR-H also did so, but it ascribed even higher weights to exchange rate variables.



Table 3. Forecast accuracy metrics

	RMSE	N-RMSE	MAE	MASE
BDMM-SS	8.9581	0.0945	6.6615	2.1327
BDMM-SS-A	4.2094	0.0444	3.1482	1.0079
BDMM-SS-H	4.4507	0.0470	3.3030	1.0575
BDMM-SS-K	8.8260	0.0931	6.6613	2.1327
BDMM-SS-A-K	3.9377	0.0416	2.9582	0.9471
BDMM-SS-H-K	4.0204	0.0424	3.1099	0.9957
BDMM-NR	4.1208	0.0435	3.0920	0.9899
BDMM-NR-H	3.8553	0.0407	2.9276	0.9373
BDMM-NR-1MOD	4.2338	0.0447	3.1337	1.0033
DMA	3.7607	0.0397	2.8458	0.9111
DMS	3.8380	0.0405	2.9448	0.9428
DMA-1VAR	3.9542	0.0417	2.9751	0.9525
DMS-1VAR	4.0874	0.0431	3.0604	0.9798
BMA	3.8074	0.0402	2.9053	0.9301
BMS	3.9057	0.0412	3.0041	0.9618
BMA-1VAR	4.0663	0.0429	3.0733	0.9839
BMS-1VAR	4.0714	0.0430	3.0697	0.9828
DMA-K	3.7462	0.0395	2.8108	0.8999
DMS-K	3.8248	0.0404	2.8649	0.9172
DMA-1VAR-K	3.9090	0.0412	2.9452	0.9429
DMS-1VAR-K	3.8988	0.0411	2.9262	0.9369
BMA-K	3.8167	0.0403	2.8613	0.9161
BMS-K	3.8460	0.0406	2.8940	0.9265
BMA-1VAR-K	4.0183	0.0424	2.9975	0.9597
BMS-1VAR-K	4.0180	0.0424	2.9963	0.9593
LASSO	3.9114	0.0413	2.9295	0.9379
RIDGE	3.8333	0.0405	2.8919	0.9259
EL-NET	3.8448	0.0406	2.9002	0.9285
B-LASSO	3.8463	0.0406	2.8938	0.9265
B-RIDGE	3.9209	0.0414	2.9417	0.9418
LARS	3.9901	0.0421	3.0171	0.9659
TVP	5.2093	0.0550	3.9573	1.2670
TVP-FOR	4.2491	0.0448	3.2028	1.0254
TVP-K	4.1956	0.0443	3.0911	0.9896
TVP-FOR-K	4.2250	0.0446	3.1473	1.0076
ARIMA	4.0074	0.0423	3.0326	0.9709
NAIVE	4.1940	0.0443	3.1235	1.0000
HA	20.1314	0.2124	15.2861	4.8939
HA-ROLL	20.1629	0.2128	15.8311	5.0684



Table 4. The Clark-West test outcomes

larger	null	CW statistic	CW p-value
BDMM-SS	TVP	1.3743	0.0847
BDMM-SS-A	TVP	5.4306	0.0000
BDMM-SS-H	TVP	5.1591	0.0000
BDMM-SS-K	TVP-K	-1.1000	0.8643
BDMM-SS-A-K	TVP-K	4.4544	0.0000
BDMM-SS-H-K	TVP-K	3.8545	0.0001
BDMM-NR	BDMM-NR-1MOD	4.5229	0.0000
BDMM-NR-H	BDMM-NR-1MOD	4.6307	0.0000
DMA	TVP-FOR	5.2999	0.0000
DMS	TVP-FOR	5.2130	0.0000
BMA	TVP	6.5274	0.0000
BMS	TVP	6.5215	0.0000
DMA-K	TVP-FOR-K	4.0591	0.0000
DMS-K	TVP-FOR-K	3.7423	0.0001
BMA-K	TVP-K	3.5581	0.0002
BMS-K	TVP-K	3.4871	0.0002

5. CONCLUSIONS

Forecasting base metal prices index with various Bayesian-based methods, focusing on Bayesian dynamic mixture models was discussed, both in original versions and with further improvements. A comprehensive large set of economic indicators was applied, and models were estimated recursively, mimicking real-time forecasting conditions. Dynamic Model Averaging was found to be the most accurate forecasting scheme, whereas out of Bayesian dynamic mixture models, the scheme with normal regression components and selection was the most accurate. The proposed methods outperformed ARIMA or no-change forecast. For robustness, models based on released data were compared with those based on revised data. According to the applied statistical tests, there is a gain in forecast accuracy from applying more advanced model combination schemes over simple models.

6. ACKNOWLEDGEMENTS

This research was funded in whole by the National Science Centre, Poland, grant number 2022/45/B/HS4/00510. For the purpose of Open Access, the authors have applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

Republishing the full data is unavailable due to copyright restrictions. The majority of the data applied in the study are openly available as cited in the text (suitable tickers provided) and in References. The forecast data generated in the study are openly available in Figshare at <https://doi.org/10.6084/m9.figshare.28382480>.



REFERENCES

- [1] H. Zhang, H. Nguyen, D.-A. Vu, X.-N. Bui, B. Pradhan, "Forecasting monthly copper price: A comparative study of various machine learning-based methods," *Resources Policy*, vol. 73, p. 102189, 2021. Available: <https://doi.org/10.1016/j.resourpol.2021.102189>
- [2] P. Pincheira Brown, N. Hardy, "Forecasting base metal prices with the Chilean exchange rate," *Resources Policy*, vol. 62, pp. 256-281, 2019. Available: <https://doi.org/10.1016/j.resourpol.2019.02.019>
- [3] T. Shi, C. Li, W. Zhang, Y. Zhang, "Forecasting on metal resource spot settlement price: New evidence from the machine learning model," *Resources Policy*, vol. 81, p. 103360, 2023. Available: <https://doi.org/10.1016/j.resourpol.2023.103360>
- [4] H. Guo, J. Wang, Z. Li, H. Lu, L. Zhang, "A non-ferrous metal price ensemble prediction system based on innovative combined kernel extreme learning machine and chaos theory," *Resources Policy*, vol. 79, p. 102975, 2022. Available: <https://doi.org/10.1016/j.resourpol.2022.102975>
- [5] H. Ben Ameur, S. Boubaker, Z. Ftiti, et al., "Forecasting commodity prices: Empirical evidence using deep learning tools," *Annals of Operations Research*, vol. 339, pp. 349-367, 2024. Available: <https://doi.org/10.1007/s10479-022-05076-6>
- [6] World Bank, "Commodity markets." Accessed Feb. 3, 2025. [Online.]: Available: <https://www.worldbank.org/en/research/commodity-markets>
- [7] I. Nagy, E. Suzdaleva, "Mixture estimation with state-space components and Markov model of switching," *Applied Mathematical Modelling*, vol. 37, pp. 9970-9984, 2013. Available: <https://doi.org/10.1016/j.apm.2013.05.038>
- [8] I. Nagy, E. Suzdaleva, M. Karny, T. Mlynarova, "Bayesian estimation of dynamic finite mixtures," *International Journal of Adaptive Control and Signal Processing*, vol. 25, pp. 765-787, 2011. Available: <https://doi.org/10.1002/acs.1239>
- [9] D. Buncic, C. Moretto, "Forecasting copper prices with dynamic averaging and selection models," *The North American Journal of Economics and Finance*, vol. 33, pp. 1-38, 2015. Available: <https://doi.org/10.1016/j.najef.2015.03.002>
- [10] ALFRED, "Archival FRED." Accessed Feb. 3, 2025. [Online.]: Available: <https://alfred.stlouisfed.org>
- [11] FRED, "Economic data." Accessed Feb. 3, 2025. [Online.]: Available: <https://fred.stlouisfed.org>
- [12] L. Kilian, "Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market," *American Economic Review*, vol. 99, pp. 1053-1069, 2009. Available: <https://doi.org/10.1257/aer.99.3.1053>
- [13] Chicago Board Options Exchange. (2025). Historical data [Online]. Available: https://www.cboe.com/tradable_products/vix/vix_historical_data
- [14] D. Caldara, M. Iacoviello, "Measuring geopolitical risk," *American Economic Review*, vol. 112, no. 4, pp. 1194-1225, 2022. Available: <https://doi.org/10.1257/aer.20191823>
- [15] M. Iacoviello. (2025). Geopolitical risk (GPR) index [Online]. Available: <https://www.matteoiacoviello.com/gpr.htm>
- [16] Stooq, "Stooq." Accessed Feb. 3, 2025. [Online.]: Available: <https://stooq.pl/index.html>
- [17] MSCI, "End of day index data search." Accessed Feb. 3, 2025. [Online.]: Available: <https://www.msci.com/end-of-day-data-search>
- [18] FXTOP, "Historical rates." Accessed Feb. 3, 2025. [Online.]: Available: <https://fxtop.com/en/historical-exchange-rates.php>
- [19] Yahoo, "Yahoo!finance." Accessed Feb. 3, 2025. [Online.]: Available: <https://finance.yahoo.com>
- [20] P. Shrivastava, R. Vidhi, "Pathway to sustainability in the mining industry: A case study of Alcoa and Rio Tinto," *Resources*, vol. 9, p. 70, 2020. Available: <https://doi.org/10.3390/resources9060070>
- [21] J. J. Barry, G. R. Matos, W. D. Menzie, "U.S. Mineral Dependence - Statistical Compilation of U.S. and World Mineral Production, Consumption, and Trade, 1990-2010," Open-File Report 2013-1184, U.S. Geological Survey: Reston, VA, 2013. Available: <https://pubs.usgs.gov/of/2013/1184>
- [22] N. E. Idoine, E. R. Raycraft, S. F. Hobbs, P. Everett, E. J. Evans, A. J. Mills, D. Currie, S. Horn, R. A. Shaw, "World Mineral Production 2018-22," British Geological Survey: Keyworth, Nottingham, 2024. Available: <https://nora.nerc.ac.uk/id/eprint/537241/1/World%20Mineral%20Production%202018%20to%202022.pdf>
- [23] D. Croushore, "Frontiers of real-time data analysis," *Journal of Economic Literature*, vol. 49, pp. 72-100, 2011. Available: <https://doi.org/10.1257/jel.49.1.72>
- [24] K. Drachal, "dynmix': An R package for the estimation of dynamic finite mixtures," *SoftwareX*, vol. 22, p. 101388, 2023. Available: <https://doi.org/10.1016/j.softx.2023.101388>
- [25] A. E. Raftery, M. Karny, P. Ettler, "Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill," *Technometrics*, vol. 52, pp. 52-66, 2010. Available: <https://doi.org/10.1198/TECH.2009.08104>
- [26] K. Drachal, "Dynamic Model Averaging in economics and finance with fDMA: A package for R," *Signals*, vol. 1, pp. 47-99, 2020. Available: <https://doi.org/10.3390/signals1010004>



- [27] J. Friedman, T. Hastie, R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, pp. 1–22, 2010. Available: <https://doi.org/10.18637/jss.v033.i01>
- [28] R. B. Gramacy. (2019). monomvn: Estimation for MVN and Student-t Data with Monotone Missingness [Online]. Available: <https://CRAN.R-project.org/package=monomvn>
- [29] T. Hastie, B. Efron. (2013). lars: Least Angle Regression, Lasso and Forward Stagewise [Online]. Available: <https://CRAN.R-project.org/package=lars>
- [30] R. J. Hyndman, Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *Journal of Statistical Software*, vol. 26, pp. 1–22, 2008. Available: <https://doi.org/10.18637/jss.v027.i03>
- [31] G. Koop, D. Korobilis, "Forecasting inflation using dynamic model averaging," *International Economic Review*, vol. 53, pp. 867–886, 2012. Available: <https://doi.org/10.1111/j.1468-2354.2012.00704.x>
- [32] R. J. Hyndman, A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, pp. 679–688, 2006. Available: <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- [33] F. X. Diebold, R. S. Mariano, "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, vol. 13, pp. 253–263, 1995. Available: <https://www.jstor.org/stable/1392155>
- [34] P. R. Hansen, A. Lunde, J. Nason, "The model confidence set," *Econometrica*, vol. 79, pp. 453–497, 2011. Available: <https://doi.org/10.3982/ECTA5771>
- [35] R. Giacomini, B. Rossi, "Forecast comparisons in unstable environments," *Journal of Applied Econometrics*, vol. 25, pp. 595–620, 2010. Available: <https://doi.org/10.1002/jae.1177>
- [36] T. E. Clark, K. D. West, "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, vol. 138, pp. 291–311, 2007. Available: <https://doi.org/10.1016/j.jeconom.2006.05.023>



SIMULATION OF ELECTRICAL POWER SYSTEMS OF GREECE (CRETE) USING THE InterPSS

Dionysia Svarna¹,
[0009-0005-1673-6084]

Alexios Serafeim Nterekas^{2*},
[0009-0006-4806-4574]

Georgia Ntereka³
[0009-0008-6108-3387]

¹Medical University of Sofia,
Faculty of Pharmacy,
Sofia, Bulgaria

²Technical University of Crete,
Management Systems Laboratory,
School of Production Engineering and
Management,
Chania, Greece

Abstract:

The electrical power system network is a highly complex and extensive entity, particularly when incorporating distributed generation, renewable energy sources, and energy storage devices. This research examines the load flow analysis of the Crete electrical grid using the Internet Technology based Power System Simulator (InterPSS). The Crete electrical network, consisting of 16 buses, was simulated under two scenarios. In the first scenario, the loads absorb 81.5% of their maximum power and 100% of the wind generation. In the second scenario, the loads remain the same, but the wind generation is multiplied by a factor of 5.4, resulting in a total of 1000MW=10 P.U.

The results show that in the first scenario, the voltages remain within permissible limits, and no power flow overloading occurs in the network branches, ensuring that the system operates without issues. However, in the second scenario, although the voltage remains within permissible limits, overloading is observed in several branches, making it possible for the system to operate under these conditions.

Keywords:

Simulation, Electrical Power System Network, Renewable Energy Sources, Load Flow Analysis, Internet Technology based Power System Simulator.

INTRODUCTION

The integration of renewable energy technologies and strategies for improving energy efficiency has become increasingly crucial to meet the ambitious target of reducing greenhouse gas emissions by 2050 [1]. To achieve a sustainable balance between energy production and consumption, several changes must still be addressed. These challenges are closely tied to the rising demand for energy [2], driven by industrial and economic growth in both developed and emerging nations [3]. Consequently, many countries have recognized the need to create and implement a new, sustainable energy framework across key sectors, including buildings, transportation, and industry. This new framework must be diverse and adaptable, incorporating principles such as the waste-to-energy approach [4], energy consumption reduction, the circular economy [5], improved quality of life for people [6], pollution mitigation and climate change adaptation [7], and the overarching goal of achieving carbon neutrality and zero emissions by 2050 [8].

Correspondence:

Alexios Serafeim Nterekas

e-mail:

anterekas@tuc.gr



Furthermore, to enhance the adoption of renewable energy in society, it is essential to carefully address issues related to the management of surplus renewable electricity generation [9]. This surplus can lead to significant challenges in terms of grid stability and management [10].

The primary objective of this research is to examine the load flow analysis of the Crete electrical grid using the Internet Technology based Power System Simulator (InterPSS). InterPSS has been used widely in research to design, analysis, and simulation of power systems [11].

2. INTERNET TECHNOLOGY BASED POWER SYSTEM SIMULATOR (INTERPSS)

The InterPSS (Internet technology based Power System Simulator) is a free and open-source software tool for the design, analysis, diagnosis, and operation of electrical power systems. It provides a flexible and extensible platform that allows researchers and developers to expand the simulation engine or adapt the platform for domain-specific or cross-domain power system simulation applications [11], [12]. InterPSS is both easy to use and highly functional. While many existing power system

simulation tools rely on outdated technologies, InterPSS distinguishes itself through its open and loosely coupled architecture. This innovative design enables components developed by other users to be easily integrated into the InterPSS environment, enhancing its functionality. At the same time, InterPSS components can be effortlessly incorporated into other software systems [11], [12]. The present research examines the load flow analysis of the Cretan power grid under two scenarios. The second scenario is simulated in the same manner as the first, with the only difference being that the wind power generation is multiplied by a factor of 5.4, resulting in a total of 1000 MW = 10 P.U. The software includes a graphical editor that allows users to create one-line diagrams and input the necessary data [11], [12]. The latest version offers essential features such as creating simple one-line diagrams of power systems, entering data, executing simulations, and generating reports. The electrical grid of Crete, consisting of 16 buses, is described below, with its topology illustrated in Figure 1 [13] (buses: 1-Chania, 2-Heraklion, 3-Atherinolakkou, 4-Sitia, 5-Agios Nikolaos, 6-Ierapetra, 7-Stalida, 8-Praitorion, 9-Moires, 10-Agias Barbaras, 11-Ammou, 12-Spili, 13-Rethumno, 14-Vrisses, 15-Aguias, 16-Kissamos). The design of the Cretan grid in InterPSS is depicted in Figure 2.

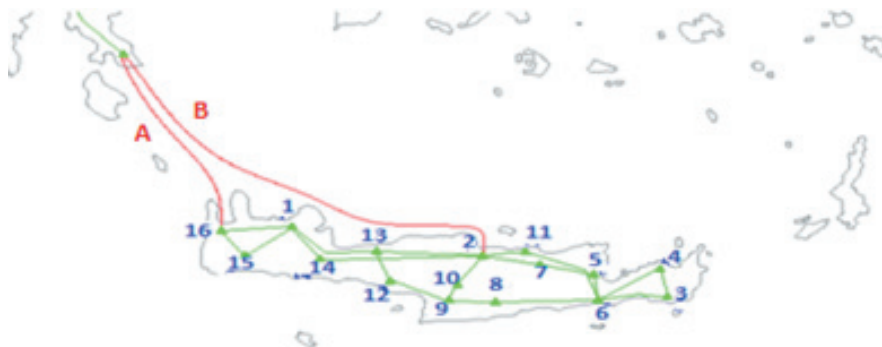


Figure 1. Buses [13]

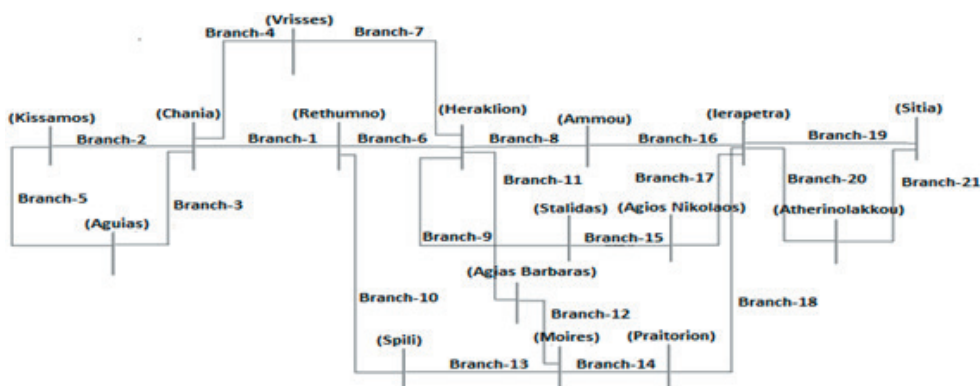


Figure 2. Electricity network of Crete in the InterPSS



3. RESEARCH METHOD

3.1. DATA ENTRY IN THE INTERPSS

The classification of buses (PV, PQ, and Swing) and the calculations of load and branch data are presented in Tables 1, and 2. Branch Data: To complete the data-set for each branch, their lengths were estimated approximately using Google Maps [11]. The data used to calculate the complex impedance of each branch are as follows: For branch type E/150: R(P.U.)= 0,000812, X(P.U.)= 0,001795, B(P.U.)= 0,00197. For branch type B/150: R(P.U.)= 0,000433, X(P.U.)= 0,001875, B(P.U.)= 0,001924. The power transmission limits applied for the 150 kV transmission lines are: MVARating1= 202 MVA, MVARating2= 124 MVA, MVARating3= 170 MVA [14]. Bus data: The calculations for each bus category are outlined below. Buses PV: Pgen (P.U.) is determined by summing all power outputs from conventional generation units connected to the same bus. The total generation is then divided by 100 MVA, which serves as the base power value [15]. Vspec (P.U.)= 1,02 for all PV and Swing buses [14]. Pload (P.U.) is divided by 100 MVA, which is the base price for power [16]. Qload (P.U.) is calculated by the type $Q_{load} = P_{load} \cdot (\sqrt{1-0.85^2} / 0.85)$ where 0.85 is the power fac-

tor ($\cos\phi$) of the load. Buses PQ: Pgen (P.U.)= +Pwind, where Pwind is from wind power production. The total production will be divided by 100 MVA which is the base price for power. Qgen (P.U.)= Qw [17], is calculated by the type $Q_w = -P_{wind} \cdot (\sqrt{1-0.85^2} / 0.85)$. Qload (P.U.) is calculated by the type $Q_{load} = P_{load} \cdot (\sqrt{1-0.85^2} / 0.85)$ where 0.85 is the power factor ($\cos\phi$) of the load [18].

In the second scenario, the simulation was conducted in the same manner as Scenario 1, with the key difference being that wind generation was multiplied by a factor of 5.4, resulting in a total power output of 1000 MW = 10 P.U. The data used in this scenario are presented below. The calculations performed for each bus are presented in Table 2.

3.2. FACE VALIDITY AND MEASURES

The face validity of the research was carefully assessed through the selection of realistic scenarios, ensuring that the results of the power flow study reflect what would be expected in a real-world power system. The scenarios chosen were based on expert input and industry standards, ensuring that the conditions modeled in the InterPSS software accurately represent typical operational states of modern power grids [14], [15].

Table 1. Calculations for each bus - First Scenario

Buses	Type	$\cos\phi$	Pgen (P.U.)	Vspec (P.U.)	Qgen (P.U.)	Pload (P.U.)	Qload (P.U.)
Chania	PV	0.85	2	1.02		1.26325	0.805228488
Heraklion	PV	0.85	1.98	1.02		0.9291	0.596110331
Atherinolakkou	PV	0.85	1.01	1.02		0.1467	0.090916494
Sitia	PQ	0.85	0.89		0.1807222	0.163	0.101018327
Agios Nikolaos	PQ	0.85	0.12		-0.024367	0.4564	0.282851316
Ierapetra	PQ	0.85	0		0	0.20375	0.126272909
Stalida	PQ	0.85	0		0	0.6357	0.393971476
Praitorion	PQ	0.85	0		0	0.13855	0.085865578
Moires	PQ	0.85	0.18		0.0365506	0.38305	0.237393069
Agias Barbaras	PQ	0.85	0.15		0.0304588	0.12225	0.075763745
Ammou	PQ	0.85	0		0	0.7498	0.464684305
Spili	PQ	0.85	0		0	0.1956	0.121221993
Rethumno	PQ	0.85	0.1		0.0203059	0.6846	0.424276974
Vrisses	PQ	0.85	0		0	0.2282	0.141425658
Aguias	PQ	0.85	0		0	0.1304	0.080814662
Kissamos	Swing	0.85	0.19	1.02	-0.0385811	0.163	0.101018327

**Table 2.** Calculations for each bus - Second scenario

Buses	Type	cos ϕ	Pgen (P.U.)	Vspec (P.U.)	Qgen (P.U.)	Pload (P.U.)	Qload (P.U.)
Chania	PV	0.85	2.4873	1.02		1.26325	0.912443461
Heraklion	PV	0.85	2.423	1.02		0.9291	0.695202957
Atherinolakkou	PV	0.85	1.01	1.02		0.1467	0.090916494
Sitia	PQ	0.85	4.8327		1.0481888	0.163	0.101018327
Agios Nikolaos	PQ	0.85	0.6516		0.1413288	0.4564	0.282851316
Ierapetra	PQ	0.85	0		0	0.20375	0.126272909
Stalida	PQ	0.85	0		0	0.6357	0.393971476
Praitorion	PQ	0.85	0		0	0.13855	0.085865578
Moires	PQ	0.85	0.9774		0.2119932	0.38305	0.237393069
Agias Barbaras	PQ	0.85	0.8145		-0.176661	0.12225	0.075763745
Ammou	PQ	0.85	0		0	0.7498	0.464684305
Spili	PQ	0.85	0		0	0.1956	0.121221993
Rethumno	PQ	0.85	0.543		-0.117774	0.6846	0.424276974
Vrisses	PQ	0.85	0		0	0.2282	0.141425658
Aguias	PQ	0.85	0		0	0.1304	0.080814662
Kissamos	Swing	0.85	1.0317	1.02	0.2237706	0.163	0.101018327

The face validity of the scenarios was also confirmed by ensuring that the system's behavior in these conditions aligned with expectations from both theoretical knowledge and real-world experience. For instance, voltage violations outside the acceptable range and overloading of transmission lines would typically signal potential issues in system stability or capacity [18]. The results of the analysis demonstrated that all simulated conditions remained within the expected operational limits, supporting the face validity of the model. These steps confirm that the research provides plausible and reasonable results that would be recognized as valid by professionals in the field.

3.3. CONTENT RELIABILITY AND VALIDITY

The reliability and validity of the results in the power flow analysis using the Newton-Raphson method are ensured through the precise application of the method within the InterPSS software, alongside adherence to established industry standards. The Newton-Raphson method, widely used for solving nonlinear equations in power systems, provides reliable results as it allows for rapid convergence and high accuracy in power flow analysis [19]. The voltages at the system buses must lie within the range of $1.05 \text{ P.U.} < V < 0.95 \text{ P.U.}$, as required for proper system operation [20]. Furthermore, the 202 MVA, 124 MVA, and 170 MVA transfer limits for the 150 kV

lines provide a safe basis for avoiding overloads and effectively distributing power throughout the network [14]. The use of InterPSS ensures that the simulation results are both valid and reliable, confirming that the network operates within technical and operational limits.

4. RESULTS

4.1. FIRST SCENARIO

The load flow analysis using the Newton-Raphson method, with a maximum of 50 iterations and a convergence threshold of 0.0001 [19]. We examine the voltage (P.U.), levels of all buses. The voltage fluctuation range is normally between $0.95 \text{ P.U.} < V < 1.05 \text{ P.U.}$ [20]. As observed, no bus voltage exceeds these limits. Therefore, no operational issues are present in this scenario (1-Chania: 1.02 P.U., 2-Heraklion: 1.02 P.U., 3-Atherinolakkou: 1.02 P.U., 4-Sitia: 1.0153 P.U., 5-Agios Nikolaos: 0.9902 P.U., 6-Ierapetra: 0.9981 P.U., 7-Stalida: 0.9903 P.U., 8-Praitorion: 0.9953 P.U., 9-Moires: 0.996 P.U., 10-Agias Barbaras: 1.0043 P.U., 11-Ammou: 0.9984 P.U., 12-Spili: 0.9926 P.U., 13-Rethumno: 0.9949 P.U., 14-Vrisses: 1.0169 P.U., 15-Aguias: 1.0188 P.U., 16-Kissamos: 1.02 P.U.). There was no issue in this scenario, as all bus voltages remained within the permissible limits (Figure 3).



Next, the transmitted power in each branch of the system was examined. If the power limits were exceeded (overloading), additional branches would be required to redistribute the power flow. The transmission power limits used for the 150 kV transmission lines are as follows MVARating1= 202 MVA, MVARating2= 124 MVA MVARating1= 170 [14]. The analysis focused on the highest transmission power limit, 202 MVA. The results, which $S = \sqrt{P^2 + Q^2}$ show the apparent power (S) for each branch, while P and Q represent the active and reactive power, respectively indicate that there was not a power flow over flow overload in several branches [14].

4.2. SECOND SCENARIO

The second scenario is simulated in the same manner as scenario 1, with the only difference being that wind power generation is multiplied by a factor of 5.4, resulting in a total of 1000MW = 10 P.U. The load flow analysis using the Newton-Raphson method, with a maximum of 50 iterations and a convergence threshold of 0.0001 [19].

We examine the voltage (P.U.), levels of all buses. The voltage fluctuation range is normally between 0.95 P.U. < V < 1.05 P.U. [20]. As observed, no bus voltage exceeds these limits. Therefore, no operational issues are present in this scenario (1-Chania: 1.02 P.U., 2-Heraklion: 1.02 P.U., 3-Atherinolakkou: 1.02 P.U., 4-Sitia: 1.0079 P.U., 5-Agios Nikolaos: 0.9552 P.U., 6-Ierapetra: 0.9612 P.U., 7-Stalida: 0.9656 P.U., 8-Praitorion:

0.9626 P.U., 9-Moires: 0.9706 P.U., 10-Agias Barbaras: 0.9903 P.U., 11-Ammou: 0.982 P.U., 12-Spili: 0.9556 P.U., 13-Rethumno: 0.9579 P.U., 14-Vrisses: 0.9961 P.U., 15-Aguias: 1.0131 P.U., 16-Kissamos: 1.02 P.U.). There was no issue in this scenario, as all bus voltages remained within the permissible limits (Figure 4).

Next, the transmitted power in each branch of the system was examined. If the power limits were exceeded (overloading), additional branches would be required to redistribute the power flow. The transmission power limits used for the 150 kV transmission lines are as follows MVARating1= 202 MVA, MVARating2= 124 MVA MVARating1= 170 [14]. The analysis focused on the highest transmission power limit, 202 MVA. The results include the apparent power (S) for each branch, calculated as , and presented in the last column. Furthermore, P and Q represent the active and reactive power, respectively, indicating that there was no power flow overload in several branches [14].

Table 3 presents the branches where current flow overloads occurred. To redistribute the power, additional branches will be required. These branches are located between the following buses: 1-13, 1-15, 1-14, 1-16, 2-14, 3-6, 6-4, 13-1, 14-1, 14-2, 15-1, 15-16, 16-1, 16-15. The corresponding power flows are presented in Table 3.

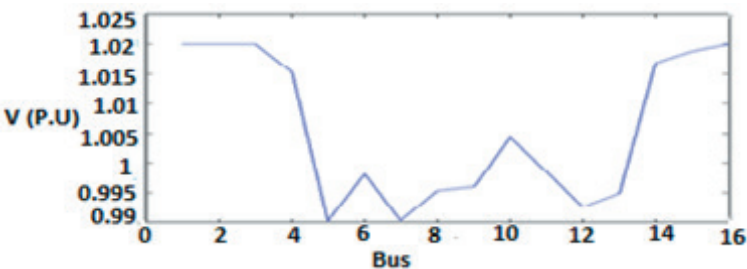


Figure 3. Voltage diagram

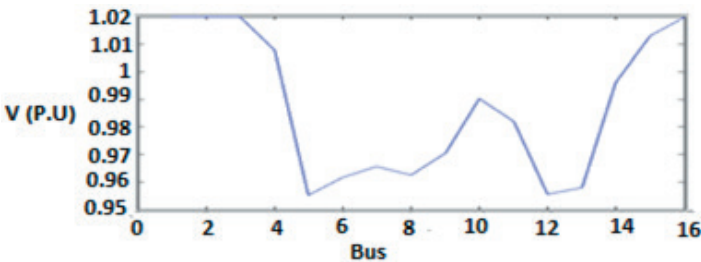


Figure 4. Voltage diagram



Table 3. Power Flows in the Branches

Previous Bus	Net Bus	P (MW)s	Q (MVAr)	$S = \sqrt{P^2 + Q^2}$ (MVA)
1	13	-280.8	178.81	332.8988677
	15	261.29	-31.22	263.148537
	14	-206.14	111.36	234.2962851
	16	347.64	-44.11	350.4272559
2	14	254.74	-2.06	254.7483291
3	6	256.91	48.08	261.3703015
	6	291.81	-34.33	293.8224379
6	4	-265.65	84.86	278.8747785
	3	-246.91	-11.38	247.1721111
	1	304.23	-86.73	316.3510484
14	1	211.67	-92.07	230.8269347
	2	-233.67	78.07	246.3667871
15	1	-252.67	62.6	260.3092179
	16	239.67	-70.6	249.852094
16	1	-330.32	112.37	348.910188
	15	-235.48	85.59	250.5523468

Table 4. Power Flows in the Branches after System Reinforcement

Previous Bus	Net Bus	P (MW)s	Q (MVAr)	$S = \sqrt{P^2 + Q^2}$ (MVA)	MVARating1
1	13	-266.2	140.09	300.8116489	404
	15	279.3	-51.61	284.0283121	404
	14	-263.19	97.5	280.6692468	404
	16	372.08	-69.32	378.4822173	404
2	14	304.22	-41.74	307.0700832	404
	11	-104.15	107.94	149.9942202	202
	7	-127.68	104.94	165.2712498	202
	10	-65.18	84.57	106.773205	202
	13	142.79	2.96	142.8206767	202
3	6	220.97	51.26	226.8376699	404
	4	-133.97	52.14	143.7586189	202
4	3	135.27	-49.5	144.0424344	202
	6	328.73	-64.5	334.9980043	404
5	6	-179.05	-1.81	179.0591483	202
	7	199.05	-40.19	203.0668328	202
6	4	-312.07	82.92	322.8984535	404
	3	-217.21	-48.37	222.5305395	404
	11	188.08	-32.21	190.8181608	202



Previous Bus	Net Bus	P (MW)s	Q (MVar)	$S = \sqrt{P^2 + Q^2}$ (MVA)	MVARating1
	5	180.8	7.15	180.9413234	202
	8	140.4	-21.5	142.0366502	202
7	5	-194.4	55.77	202.2415707	202
	2	131.4	-94.77	162.0102247	202
8	6	-135.96	31.4	139.5388175	202
	9	122.96	-39.4	129.1182466	202
9	8	-121.4	42.23	128.5353372	202
	10	-1.62	-62.05	62.07114386	202
	12	182.02	-24.18	183.6190426	202
10	9	1.99	60.53	60.56270304	202
	2	67.01	-84.53	107.8687211	202
11	6	-180.06	57.56	189.0363912	202
	2	106.06	-103.56	148.2342646	202
12	9	-177.54	37.98	181.5569663	202
	13	158.54	-49.98	166.2315614	202
13	2	-138.19	6.46	138.3409112	202
	12	-155.71	57.97	166.1509103	202
	1	275.9	-117.43	299.8509878	404
14	1	267.16	-90.15	281.9600825	404
	2	-289.16	76.15	299.0189427	404
15	1	-274.38	61.25	281.1333258	404
	16	261.38	-69.25	270.3979787	404
16	1	-362.1	99.22	375.4477572	404
	15	-258.99	73.54	269.2284378	404

In addition, power flow overloads were observed in the branches, resulting in the network being unable to operate under these conditions. Particularly in the branches between the buses 1-13 (332.9 MVA), 16-1(350.4 MVA), 1-15(263.1 MVA), 1-14(234.3 MVA), 16-15(250.6 MVA), 2-14(254.7 MVA), 6-4(293.8 MVA), and 6-3(261.7 MVA), there was an overload exceeding MVARating1, and a second identical circuit was added. This was achieved by halving the resistance values R (P.U.) and the inductive reactance values X (P.U.), and by doubling the transverse conductance $1/2B$ (P.U.) and the power transfer limits MVARating1, MVARating2, MVARating3 [14]. The results of the network after these changes to enhance the system are presented in Table 4.

Table 4 presents the power flows in each branch of the system after the reinforcement measures were implemented. These measures were carried out by adding

additional circuits and adjusting the system parameters, such as halving the resistance and reactance values and doubling the power transfer limits, as described earlier. The table shows the updated active and reactive power flows in each branch following these changes to relieve overloading and improve system stability. After the necessary changes were made, it was observed that the branches are now within the permissible limits and are no longer overloaded. As a result, the system is functioning without any issues.



5. DISCUSSION

The results obtained from the InterPSS simulations confirm its effectiveness in analyzing power system performance. In the first scenario, where load values were maintained near their current levels with 100% of the existing wild power generation, voltage levels remained within permissible limits, and no power flow overloads were observed. This indicates that under these conditions, the system can operate reliably without modifications. In comparison, Kumar et al. [21] demonstrated that InterPSS provides results comparable to other software for flow analysis while offering a more user-friendly interface. Similarly, Zhou and Huang [11] highlighted the software's efficiency in real-time simulations, making it strong software.

In the second scenario, where wind power generation was increased to 1000 MW while keeping load levels constant, significant power overloads were detected in multiple network branches. Although voltage levels remained within acceptable limits, the increased wind penetration caused congestion in several transmission lines, rendering the system unable to function under these conditions. The findings align with Brahmendra Kumar et al. [22], who emphasized the necessity of grid reinforcements when integrating large-scale renewable energy sources. Their study confirmed that high wind power energy generation can lead to transmission congestion, a result also observed in the present study. Additionally, Medina et al. [23] examined transmission congestion in renewable-dominated grids and recommended expanding transmission capacity as an effective solution, further validating the approach proposed in this study.

Finally, the proposed solution involved adding new transmission lines between overloaded buses, thereby redistributing power flow and mitigating the overloading issue, ultimately allowing stable system operation which is in line with the findings of Wang and Li [24].

6. CONCLUSIONS

The research demonstrated that while InterPSS is a robust tool for power system analysis, grid stability under high renewable energy penetration remains a critical challenge. The results emphasize that without proper infrastructure reinforcement, increased wind power integration can lead to significant operational constraints. This highlights the urgent need for strategic transmission expansions to ensure a resilient and reliable power grid.

7. ACKNOWLEDGMENT

The authors would like to acknowledge the support at the Technical University of Crete, which facilitated this research. Appreciation is extended to colleagues and peers for their valuable discussions and insights throughout the study. Additionally, gratitude is expressed to the developers of InterPSS for making an open-source platform available for comprehensive power system analysis.

REFERENCES

- [1] P. Tsvetkov, P. Samuseva, L. Nikolaychuk, "The research of the impact of energy efficiency on mitigating greenhouse gas emissions at the national level," *Energy Conversion and Management*, vol. 314, p. 118671, 2024, <https://doi.org/10.1016/j.enconman.2024.118671>.
- [2] K. Devadasa, N. H. Laxminarayana, "The role of renewable energy in mitigating climate change," *Journal of Visual and Performing Arts*, 2024, <https://doi.org/10.29121/shodhkosh.v4.i1.2023.2786>.
- [3] F. S. Chien, K. Y. Chau, M. Sadiq, G. L. Diep, T. K. Tran, T. H. A. Pham, "What role renewable energy consumption, renewable electricity, energy use and import play in environmental quality?," *Energy report*, vol. 10, pp. 3826-3834, Nov. 2023, <https://doi.org/10.1016/j.egy.2023.10.024>.
- [4] A. Ramos, "Sustainability assessment in waste management: An exploratory study of the social perspective in waste-to-energy cases," *Journal of Cleaner Production*, vol. 475, 2024, <https://doi.org/10.1016/j.jclepro.2024.143693>.
- [5] M. Busu, C. L. Trica, "Sustainability of Circular Economy Indicators and Their Impact on Economic Growth of the European Union," *Sustainability*, vol. 11, no. 19, p. 5481, 2019, <https://doi.org/10.3390/su11195481>.
- [6] J. Rosak-Szyrocka, A. A. S. Allahham, J. Zywiotek, J. A. Turi, A. Das, "Expectations for renewable energy, and its impacts on quality of life in European union countries," *Management Systems In Production Engineering*, vol. 3, no.2, pp. 128-137, 2023, <https://doi.org/10.2478/mspe-2023-0015>.
- [7] F. Wang, J. D. Harindintwali, K. Wei, et al., "Climate change: Strategies for mitigation and adaptation," *The Innovation Geoscience*, vol. 1, no. 1, p. 100015, 2023, <https://doi.org/10.59717/j.xinn-geo.2023.100015>.
- [8] L. Chen, G. Msigwa, M. Yang, A. Osman, S. Fawzy, D. W. Rooney, P-S. Yap, "Strategies to achieve a carbon neutral society: a review," *Environ Chem Lett*, vol. 20, pp. 2277-2310, 2022, <https://doi.org/10.1007/s10311-022-01435-8>.



- [9] T. Z. Ang, M. Salem, M. Kamarol, H. S. Das, M. A. Nazari, N. Prabakaran, "A comprehensive study of renewable energy sources: Classifications, challenges and suggestions," *Energy Strategy Reviews*, vol. 43, 2022, <https://doi.org/10.1016/j.esr.2022.100939>.
- [10] M. A. V. Rad, A. Kasaeian, X. Niu, K. Zang, O. Mahian, "Excess electricity problem in off-grid hybrid renewable energy systems: A comprehensive review from challenges to prevalent solutions," *Renewable Energy*, vol. 212, pp. 538-560, 2023, <https://doi.org/10.1016/j.renene.2023.05.073>.
- [11] M. Zhou, Q. Huang, "InterPSS A New Generation Power System Simulation Engine," *Engineering, Computer Science*, 2017, <https://doi.org/10.48550/arXiv.1711.10875>.
- [12] InterPSS Development Team, "InterPSS Community," [Online]. Available: <https://sites.google.com/a/interpss.org/interpss/Home>. [Accessed: Mar. 18, 2025].
- [13] E. Loukarakis, "Isolated Electrical Power Systems Optimal Interconnections Planning," Diploma Thesis, Technical University of Crete, Chania, 2012.
- [14] P. Kundur, "Power System Stability and Control," McGraw-Hill, 1994.
- [15] J. D. Glover, M. S. Sarma, T. J. Overbye, "Power System Analysis and Design," 5th ed., Cengage Learning, 2011.
- [16] A. J. Wood, B. F. Wollenberg, "Power Generation, Operation, and Control," 2nd ed., Wiley, 2012.
- [17] R. Billinton, R. N. Allan, "Reliability Evaluation of Power Systems," Springer, 1996.
- [18] H. Saadat, "Power System Analysis," 3rd ed. McGraw-Hill, 2010.
- [19] A. Ahmadi, M. C. Smith, E. R. Collins, V. Dargahi, S. Jin, "Fast Newton-Raphson Power Flow Analysis Based on Sparse Techniques and Parallel Processing," in *IEEE Transactions on Power Systems*, vol. 37, no. 3, pp. 1695-1705, 2022, <https://doi.org/10.1109/TPWRS.2021.3116182>.
- [20] K. Chaiyaphun, P. Santiprapan, K. Areerak, "A Modified Variable Power Angle Control for Unified Power Quality Conditioner in a Distorted Utility Source," *Energies*, vol. 17, no. 12, p. 2830, 2024, <https://doi.org/10.3390/en17122830>.
- [21] Y. Kumar, V. K. Devabhaktuni, S. Vemuru, "Comparison of power system simulation tools with load flow study cases," *Institute of Electrical and Electronics Engineers*, 2015, <https://doi.org/10.1109/EIT.2015.7293355>.
- [22] G. V. Brahmendra Kumar, K. Ratnam, P. Kaliannan, S. Padmanaban, J. B. Holm-Nielsen, "Large Scale Renewable Energy Integration: Issues and Solutions," *Energies*, vol. 12, no. 10, 2019, <https://doi.org/10.3390/en12101996>.
- [23] C. Medina, G. G. Gonzales, "Transmission Grids to Foster High Penetration of Large-Scale Variable Renewable Energy Sources - A Review of Challenges, Problems, and Solutions," *International Journal of Renewable Energy Technology*, vol. 12, no. 1, pp. 146-169, Jan. 2022, <https://doi.org/10.20508/ijrer.v12i1.12738.g8400>.
- [24] H. Wang, Z. Li, "A Review of Power System Transient Stability Analysis and Assessment," Prognostics and System Health Management Conference (PHM-Qingdao), China, 2019, <https://doi.org/10.1109/PHM-Qingdao46334.2019.8942834>.



MANDIBLE DOSE COMPARISON BETWEEN FOTELP-VOX SIMULATIONS AND IMRT IN HEAD AND NECK RADIOTHERAPY

Milena Živković^{1*},
[0000-0001-8567-7050]

Marina Svičević¹,
[0000-0003-2791-3849]

Dragana Krstić¹,
[0000-0002-3517-9210]

Taha Yaseen Wais²,
[0000-0003-1069-7588]

Lazar Krstić¹
[0000-0002-4703-2291]

¹University of Kragujevac,
Faculty of Science,
Kragujevac, Serbia

²University of Mosul,
Chemical, Biological and Radiological
Safety and Security Division,
Mosul, Iraq

Abstract:

Determining the absorbed dose of scattered radiation in specific body organs and tissues during radiotherapy is crucial for minimizing potential damage. This study investigates whether the FOTELP-VOX Monte Carlo simulation software can accurately estimate the absorbed dose in the mandible of patients undergoing head and neck radiotherapy, and how closely these estimates align with expert-verified IMRT treatment plans. As a representative clinical scenario, the analysis focused on ten patients with parotid gland tumors, whose treatment plans were generated using the IMRT planning system at the Clinical Center. The dosimetric parameters analyzed include minimum dose, maximum dose, and mean dose to the mandible. The average mandibular volume was $60 \pm 15 \text{ cm}^3$. Dose differences between FOTELP-VOX simulations and IMRT ranged from 5% to 8%, indicating good agreement even without prior expert adjustment. These results suggest that the absorbed dose to the mandible remains within acceptable limits, although careful planning remains essential to minimizing unnecessary exposure. The findings support the potential use of FOTELP-VOX as a supplementary dosimetric tool for quality assurance, particularly in estimating scattered dose to critical structures adjacent to the target volume. The simulation outcomes confirmed the software's capability to generate accurate three-dimensional dose distributions resulting from particle interactions within complex anatomical structures. Further improvements and clinical validation are necessary to enhance the robustness and clinical integration of this approach.

Keywords:

Head and Neck Radiotherapy, Mandible Dose Estimation, Monte Carlo Simulation, FOTELP-VOX, IMRT Technique.

INTRODUCTION

Radiotherapy for head and neck cancers requires precise dose delivery to the tumor while minimizing exposure to surrounding healthy tissues and critical organs. Among the most sensitive structures during such treatments are the salivary glands (parotid, submandibular, and sublingual glands), which are particularly vulnerable to radiation-induced damage. In 2022, their global incidence ranked 28th, with 55,083 cases, and their mortality ranked 27th, with 23,942 deaths [1]. The parotid glands, located in front of the ears, are the largest salivary glands, responsible for saliva production and oral health.

Correspondence:

Milena Živković

e-mail:

milena.zivkovic@pmf.kg.ac.rs





Excessive radiation to these glands can lead to xerostomia (dry mouth), significantly affecting the patient's quality of life. Other important organs at risk (OARs) include the mandible, spinal cord, brainstem, optic nerves, and eyes. The protection of these structures is crucial during treatment planning to avoid complications such as osteoradionecrosis, especially in the case of the mandible [2]. Diagnostic procedures typically include physical examination, biopsy, and imaging techniques such as ultrasound, MRI, and CT, which assist oncologists in determining tumor size, location, and spread. Treatment strategies often involve surgical resection followed by radiation therapy, or radiotherapy alone when surgery is not feasible. Radiotherapy relies on high-energy photons, electrons, or other particles to destroy cancerous cells. For the parotid glands, the currently accepted dose threshold is around 26 Gy, although some studies indicate that higher doses may still allow partial functional preservation [3], while others propose lowering this threshold to 22.5 Gy to reduce the risk of irreversible damage [4]. While much attention has been given to direct dose delivery to the target and major glands, less focus has been placed on the impact of scattered radiation on adjacent healthy tissues—such as the mandible—which may still receive clinically relevant doses even if not directly targeted.

Modern radiation therapy relies on advanced planning systems to deliver high doses to tumors while sparing healthy tissue. Among these, Intensity-Modulated Radiation Therapy (IMRT) is one of the most widely used and clinically validated techniques. IMRT enables modulation of beam intensity from multiple directions, resulting in highly conformal dose distributions that significantly improve the sparing of critical structures compared to conventional approaches [5]. Using inverse planning algorithms and computer-controlled multileaf collimators, this technique allows for precise shaping of the dose to match the three-dimensional geometry of the tumor. Clinical studies have consistently confirmed its advantages in head and neck treatments, where IMRT has led to a notable reduction in the incidence of xerostomia and a decreased risk of mandibular osteoradionecrosis [6].

In addition to clinically implemented treatment planning systems, Monte Carlo (MC) simulations have become an essential tool in radiation dosimetry research due to their unmatched accuracy. Unlike conventional dose calculation algorithms based on approximations, MC methods simulate the physical interactions of individual particles as they travel through heterogeneous anatomical structures, taking into account effects such

as Compton scattering, the photoelectric effect, and pair production. This high level of precision makes them the gold standard for dose estimation in radiotherapy, particularly in anatomically complex or heterogeneous regions. While the computational demands of MC simulations have traditionally limited their use in routine clinical workflows, they are increasingly employed for independent verification, research applications, and special clinical cases. Well-known general-purpose MC codes such as MCNP [7] [8], PHITS [9] [10], GEANT4 [11], PENELOPE [12], and EGSnrc [13] are widely used in medical physics for modeling radiation transport and dose distribution. A specialized tool within this family is FOTELP-VOX [14] (Photon, Electron, and Positron Monte Carlo Transport Simulation – Voxel-Based), a voxel-based extension of the general-purpose FOTELP code [15]. FOTELP-VOX enables detailed simulation of radiation transport through anatomically realistic patient models derived from CT imaging. This allows researchers to obtain accurate three-dimensional dose distributions in both tumors and surrounding tissues. Previous studies have shown that FOTELP-VOX can replicate clinical dose distributions in various anatomical sites with deviations typically ranging from 5% to 8% compared to commercial treatment planning systems [16] [17]. However, while the focus of most FOTELP-VOX applications has been on tumor dosimetry, the estimation of absorbed dose in surrounding healthy tissues, such as the mandible, has not been sufficiently explored. This is especially important given that scattered radiation and beam spillover can lead to non-negligible dose deposition in nearby structures, even when they are not the direct target of treatment.

In clinical practice, IMRT planning systems are used by experienced radiation oncologists and physicists to generate optimized dose distributions that meet therapeutic objectives while sparing healthy tissue. In this study, we investigate whether the FOTELP-VOX simulation tool can independently provide comparable estimations of absorbed dose—without requiring prior expert tuning or adjustment. As a case study within the broader context of head and neck radiotherapy, we analyzed patients diagnosed with parotid gland tumors, whose treatment plans were developed using the IMRT technique at a clinical center in Kragujevac. The mandible was contoured and analyzed in both approaches, focusing on key parameters such as minimum, maximum, and mean dose. This comparison serves not only to evaluate the accuracy of FOTELP-VOX in estimating dose to a critical organ at risk, but also to explore its potential applicability in practice—as a complementary Monte



Carlo-based tool for quality assurance and secondary verification, particularly when assessing scattered dose to structures outside the primary target volume. By demonstrating how closely FOTELP-VOX results align with expert-verified IMRT plans, even when used in a straightforward, non-specialist setup, this study contributes to bridging the gap between research-level simulations and practical clinical implementation. It highlights the feasibility of using high-fidelity voxel-based Monte Carlo simulations as a reliable supplement to conventional planning systems for evaluating dose to sensitive anatomical structures such as the mandible.

2. MATERIALS AND METHODS

This section provides an overview of the clinical data and simulation procedures used in this study. The analysis is based on real patient cases involving head and neck cancer treatment, with radiotherapy plans created using the IMRT system at the Clinical Center Kragujevac. The simulation workflow was performed using the FOTELP-VOX Monte Carlo software, with a focus on evaluating the absorbed dose in the mandible. The following subsections describe the origin and interpretation of the patient data, as well as the simulation process and analysis of its output.

2.1. PATIENTS AND TREATMENT DATA

The study included patients diagnosed with parotid gland tumors, selected as a representative case study within the broader context of head and neck radiotherapy. These patients were treated at the University Clinical Center in Kragujevac using the IMRT technique between February 2011 and June 2021. Basic demographic information, such as age and gender distribution, is presented in Table 1. The patient group was balanced by sex (50% males and 50% females), with a mean age of 68 years.

Patient immobilization was a critical step for ensuring precise dose delivery. In cases where anatomical irregularities required it, a custom-made bolus was used, allowing optimal adaptation to the skin surface and improving dose homogeneity in the target region. Following immobilization, CT imaging was performed using a GE Discovery scanner, with a slice thickness of 2.5 mm, which ensured sufficient resolution for accurate contouring of target volumes. The delineated volumes included the Gross Tumor Volume (GTV), Clinical Target Volume (CTV), Internal Target Volume (ITV), and Planning Target Volume (PTV). Radiotherapy planning was carried out using the Varian Eclipse software (version 15.6), which provided advanced tools for optimizing treatment delivery. The Photon Optimizer algorithm was applied to adapt the dose distribution according to anatomical variations, while the Anisotropic Analytical Algorithm (AAA), version 15.6.06, was used for final dose calculation, incorporating tissue heterogeneity. Figure 1 shows the three-dimensional dose distribution and isodose contours obtained during IMRT planning.

The primary therapeutic dose was directed to the CTV, which received a mean dose of 54 Gy (range: 50–56 Gy) in daily fractions of 1.8 to 2.0 Gy. The average equivalent dose in 2 Gy fractions (EQD2) was 50 Gy. CTV1 was adequately covered by a 90% isodose line, ensuring sufficient coverage of the tumor tissue. According to published clinical dose constraints (Brouwer et al., 2015), the maximum recommended dose for the mandible is 65 Gy, while the mean dose should remain below 60 Gy. Additionally, the volume of the mandible receiving 70 Gy (V70) should not exceed 1 cm³, in order to reduce the risk of radiation-induced complications such as osteoradionecrosis.

Table 1. Characteristics of patients diagnosed with parotid gland tumors

Characteristic	Description	Value
Age (years)	Range	56-80
	Mean Age	68
Gender	Number of Males	5
	Number of Females	5
Tumor Dose	Primary	70 Gy (35 fractions)

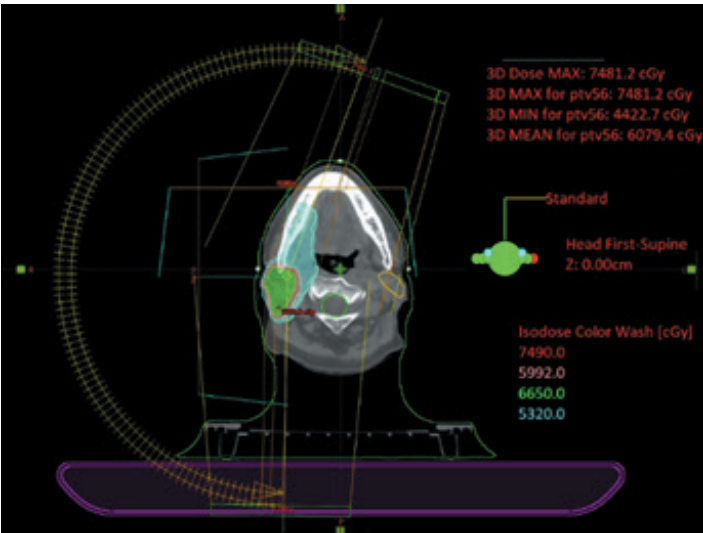


Figure 1. Three-dimensional dose distribution and isodose contours for the radiotherapy treatment

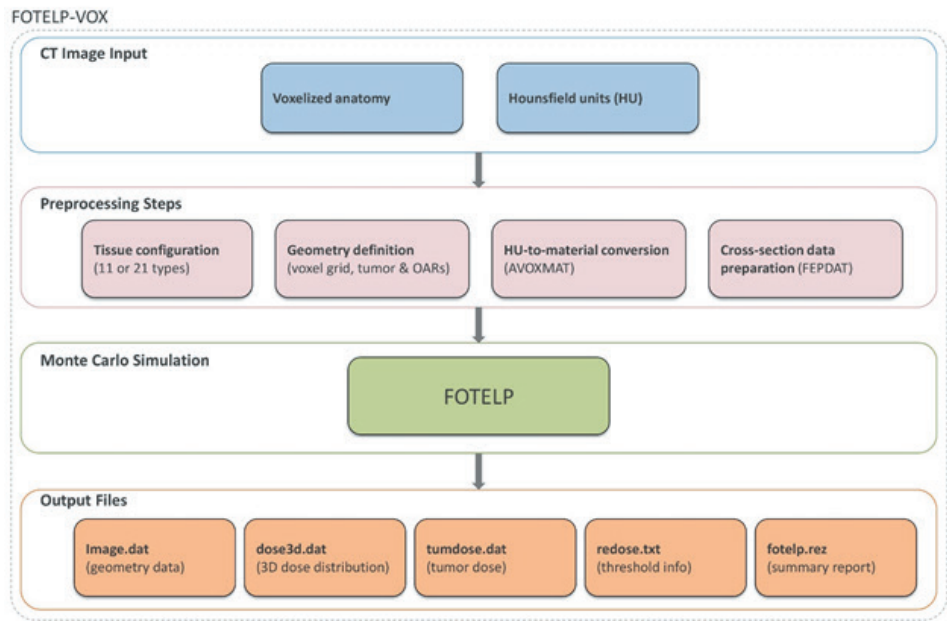


Figure 2. Overview of the FOTELP-VOX simulation workflow. The process includes CT-based voxelized geometry acquisition, preprocessing of tissue and geometry data, Monte Carlo simulation of particle transport, and generation of multiple output files for dosimetric analysis

2.2. FOTELP-VOX SIMULATION PROCESS

The FOTELP-VOX simulation process enables high-precision Monte Carlo modeling of particle transport and dose deposition based on anatomically realistic, CT-derived patient data. The workflow consists of four primary stages: CT image acquisition and voxelization, data preprocessing, FOTELP Monte Carlo simulation, and output analysis. The complete sequence is shown in Figure 2.

The process begins with importing CT images, which are converted into a voxelized representation of the patient's anatomy. Each voxel is assigned a value according to its Hounsfield number (HU), reflecting the radiological density of the tissue at that location [18]. These HU values are then used to infer material composition and tissue properties.



During preprocessing, the user selects a predefined tissue configuration file, containing either 11 or 21 material types, depending on the desired level of anatomical detail. The CT-derived voxel geometry is refined by defining a rectangular tissue region that encloses both the tumor and surrounding organs-at-risk (OARs), while minimizing the inclusion of surrounding air to improve computational efficiency. The AVOXMAT module converts HU values into material indices using a mapping formula such as:

$$MGn(i,j,k) = 10000\rho_n MH(i,j,k) \quad (1)$$

where ρ_n is the density of a given tissue type, and $MH(i,j,k)$ is the Hounsfield number at a given voxel position (i,j,k) . This process results in a material grid $MGn(i,j,k)$ that allows for high-fidelity modeling of heterogeneous tissue structures. Even small variations in HU values are preserved through unique material identifiers, enhancing the precision of dose estimation. The FEPDAT utility then prepares interaction cross-section data for photons and electrons, based on the defined tissue materials. These datasets are essential for the Monte Carlo simulation to accurately model physical processes such as Compton scattering, pair production, and photoelectric effect.

The third stage involves the execution of the Monte Carlo simulation using the FOTELP engine. It simulates the transport of photons or electrons through the voxelized patient model, iteratively computing their interactions and energy deposition in each voxel. The simulation continues until the statistical uncertainty of the dose values reaches acceptable thresholds. This enables high-resolution estimation of absorbed dose distributions throughout both target volumes and surrounding normal tissues.

In the final stage, FOTELP-VOX produces several output files essential for post-processing and clinical analysis. The *Image.dat* file contains geometric information used to reconstruct anatomical images. The *dose3d.dat* file stores a detailed three-dimensional dose distribution across the voxel grid, while *tumdose.dat* focuses specifically on the tumor region. The *redose.txt* file highlights voxels where the absorbed dose exceeds predefined safety limits, assisting in the identification of potential radiation hazards. The *fotelp.rez* file serves as a comprehensive summary, compiling simulation parameters, dose statistics, and verification data. These outputs collectively enable detailed analysis of absorbed dose distributions in both tumor and surrounding tissues, providing a foundation for further evaluation of treatment accuracy and safety.

3. RESULTS AND DISCUSSION

This study presents a dosimetric comparison between the FOTELP-VOX simulation and the clinical IMRT planning system for ten patients diagnosed with parotid gland tumors. The objective was to evaluate the agreement between these two approaches in estimating the absorbed dose in the mandible — a critical organ at risk in head and neck radiotherapy. For each patient, three key dosimetric parameters were analyzed: minimum dose (*MinD*), maximum dose (*MaxD*), and mean dose (*MeanD*) within the mandibular volume.

The results of the comparison are visualized in Figure 3, which presents individual patient data for each dosimetric parameter across both IMRT and FOTELP-VOX approaches. From the graphs, it is evident that the overall trends in dose distribution are consistent between the two techniques. FOTELP-VOX approximated the IMRT dose values with high similarity, especially for mean and minimum doses. Importantly, the maximum dose values estimated by FOTELP-VOX were consistently lower than those calculated by IMRT for all patients, with absolute differences typically ranging between 200 and 300 cGy. The observed deviations across all dosimetric parameters fall within the range of 5% to 8%, which is considered clinically acceptable for Monte Carlo-based methods.

Such agreement indicates that FOTELP-VOX, even when applied without prior tuning or expert intervention, can reliably replicate the clinically verified IMRT dose distributions. This reinforces its potential as an independent simulation tool for treatment plan verification or as a secondary check for critical organ dosimetry. The voxel-based modeling in FOTELP-VOX may also offer enhanced spatial resolution in three-dimensional dose distribution, particularly in anatomically complex regions such as the mandible.

From a clinical perspective, the observed differences are not expected to impact treatment safety, as all simulated doses remained below the established critical threshold of 65 Gy for the mandible, which is associated with the risk of osteoradionecrosis [2]. Nevertheless, precise dose estimation remains vital in treatment planning, especially for head and neck cases, where scattered radiation and out-of-field dose can affect surrounding bone and soft tissues. The ability of FOTELP-VOX to predict such effects with clinically acceptable accuracy supports its application in improving quality assurance protocols and in research focused on minimizing exposure to organs at risk.

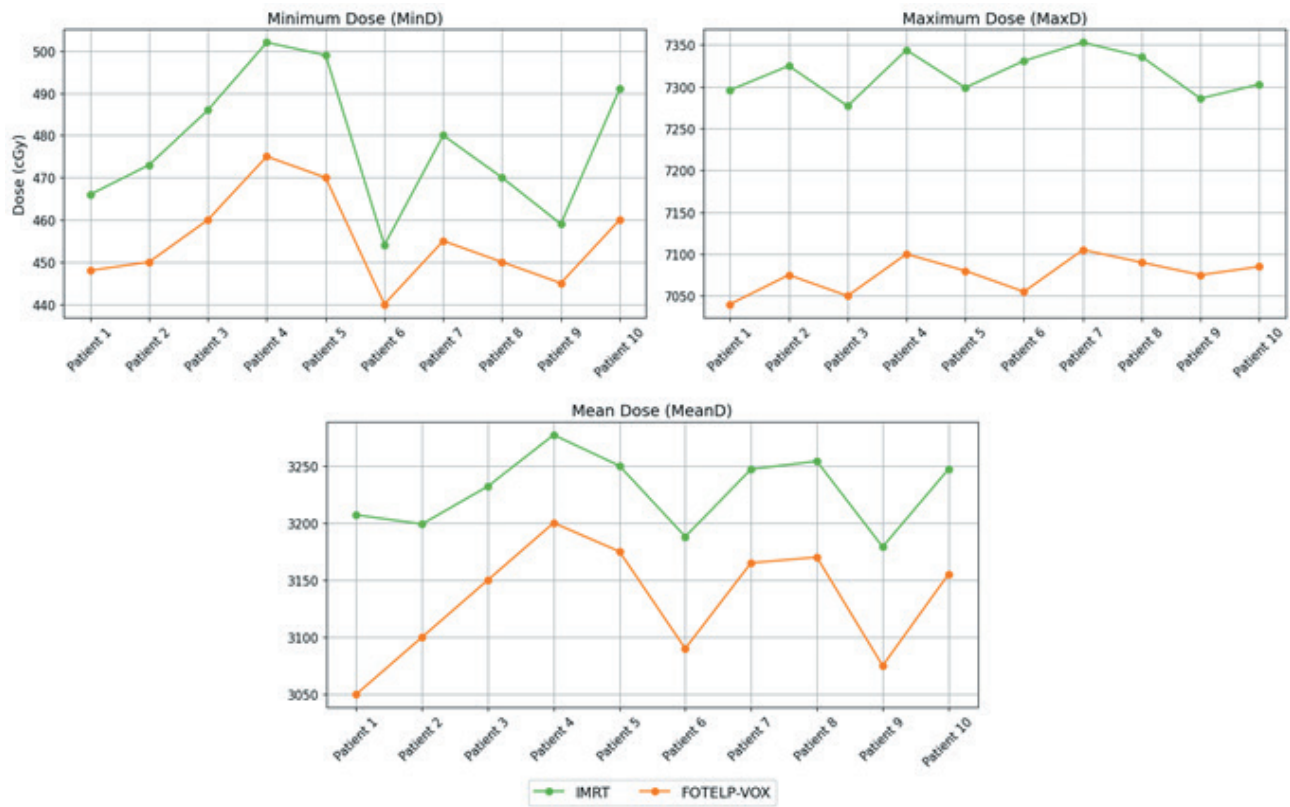


Figure 3. Comparison of IMRT and FOTELP-VOX dosimetric parameters per patient

4. CONCLUSION

This study evaluated the performance of the FOTELP-VOX Monte Carlo simulation software in estimating the absorbed dose to the mandible in head and neck radiotherapy. Using a case study of ten patients with parotid gland tumors, the absorbed dose distributions obtained with FOTELP-VOX were compared to clinically verified IMRT treatment plans. Across all analyzed patients, the simulation demonstrated a high degree of consistency with IMRT in terms of minimum, mean, and maximum dose to the mandible, with observed differences generally within 5–8%. These results confirm that FOTELP-VOX, even when used without prior expert calibration, can provide clinically relevant dose estimates for structures adjacent to the target volume. Due to its voxel-based modeling and use of patient-specific CT data, FOTELP-VOX offers high-resolution three-dimensional dose distributions, making it a promising tool for supplemental dose verification and quality assurance in radiotherapy planning. While the current application focused on the mandible as a representative organ at risk, the methodology may be extended to other critical structures affected by scattered and out-of-field radiation.

To ensure broader clinical adoption, further clinical validation of the FOTELP-VOX software across diverse anatomical scenarios is warranted. Future work could also focus on improving the usability and automation of the simulation process — for example, by incorporating automatic segmentation of tumors and organs at risk from imaging data. Additionally, tailoring simulation parameters dynamically based on the prescribed dose and anatomical complexity may further enhance the reliability and efficiency of FOTELP-VOX. These developments would facilitate its integration into clinical workflows and support more personalized and accurate treatment planning in modern radiotherapy.

5. ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia through the Agreements No. 451-03-66/2024-03/200122 and No. 451-03-65/2024-03/200122.



REFERENCES

- [1] G. C. Observatory, "Cancer statistics reports," 2024. [Online]. Available: <https://gco.iarc.who.int/en>. [Accessed 10 March 2025].
- [2] C. P. Raaijmakers, J. M. Roesink, T. Dijkema, A. C. Houweling and C. H. Terhaard, "Parotid gland function after radiotherapy," *The Lancet Oncology*, vol. 11, no. 5, p. 412, 2010.
- [3] T. Dijkema, C. P. Raaijmakers, R. K. Ten Haken, J. M. Roesink, P. M. Braam, A. C. Houweling, M. A. Moerland, A. Eisbruch and C. H. Terhaard, "Parotid gland function after radiotherapy: The combined Michigan and Utrecht experience," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 78, p. 449–453, 2010.
- [4] B. Bussels, A. Maes, P. Flamen, P. Lambin, K. Erven, R. Hermans, S. Nuyts, C. Weltens, S. Cecere and E. Lesaffre, "Dose–response relationships within the parotid gland after radiotherapy for head and neck cancer," *Radiotherapy and Oncology*, vol. 73, no. 3, p. 297–306, 2004.
- [5] RadiologyInfo.org, "Intensity-Modulated Radiation Therapy (IMRT)," [Online]. Available: <https://www.radiologyinfo.org/en/info/imrt>. [Accessed 10 March 2025].
- [6] P.-H. Mackeprang, K. Bryjova, A. E. Heusel, D. Henzen, M. Scricciolo, O. Elicin and E. H. & N. Y. I. Group, "Consideration of image guidance in patterns of failure analyses of intensity-modulated radiotherapy for head and neck cancer: a systematic review," *Radiation Oncology*, vol. 19, no. 1, p. 30, 2024.
- [7] S. A. Kalospyros, V. Gika, Z. Nikitaki, A. Kalamara, I. Kyriakou, D. Emfietzoglou, M. Kokkoris and A. G. Georgakilas, "Monte Carlo simulation-based calculations of complex DNA damage...", *Applied Sciences*, vol. 11, no. 19, p. 8985, 2021.
- [8] D. Krstic, D. Nikezic, M. Z. Jeremic, E. Dolicanin, T. B. Miladinovic and M. Zivkovic, "Comparison between MCNP and planning system in brachytherapy of cervical cancer," *Applied Radiation and Isotopes*, vol. 192, p. 110614, 2023.
- [9] Y. Gokcekuyu, F. Ekinici, A. Buyuksungur, M. S. Guzel, K. Acici and T. Asuroglu, "Comparison of X-Ray Absorption in Mandibular Tissues and Tissue-Equivalent Polymeric Materials Using PHITS Monte Carlo Simulations," *Applied Sciences*, vol. 14, no. 23, p. 10879, 2024.
- [10] R. Kakino, N. Hu, H. Tanaka, S. Takeno, T. Aihara, K. Nihei and K. Ono, "Out-of-field dosimetry using a validated PHITS model and computational phantom in clinical BNCT," *Medical Physics*, vol. 51, no. 2, pp. 1351–1363, 2024.
- [11] F. Chappuis, H. N. Tran, S. A. Zein, C. Bailat, S. Incerti, F. Bochud and L. Desorgher, "The general-purpose Geant4 Monte Carlo toolkit and its Geant4-DNA extension to investigate mechanisms underlying the FLASH effect in radiotherapy: Current status and challenges," *Physica Medica*, vol. 110, p. 102601, 2023.
- [12] O. Croce, S. Hachem, E. Franchisseur, S. Marcie, J.-P. Gerard and J.-M. Bordy, "Contact radiotherapy using a 50 kV X-ray system: Evaluation of relative dose distribution with the Monte Carlo code PENNELOPE and comparison with measurements," *Radiation Physics and Chemistry*, vol. 81, no. 6, p. 609–617, 2012.
- [13] B. A. Faddegon, I. Kawrakow, Y. Kubyshev, J. Perl, J. Sempau and L. Urban, "The accuracy of EGSnrc, Geant4 and PENELOPE Monte Carlo systems for the simulation of electron scatter in external beam radiotherapy," *Physics in Medicine & Biology*, vol. 54, no. 20, p. 6151, 2009.
- [14] M. P. Živković, T. B. Miladinović, Ž. M. Cimbalević, M. E. A. Aichouche, B. A. Pirković and D. Ž. Krstić, "FOTELP-VOX 2024: Comprehensive overview of its capabilities and applications," *Nuclear Technology and Radiation Protection*, vol. 39, no. 3, p. 212–219, 2024.
- [15] R. D. Ilic, "FOTELP-2K6, Photons, Electrons and Positrons Transport in 3D by Monte Carlo Techniques," NEA, 2007.
- [16] M. Živković, T. B. Miladinović, D. Krstić, N. Milosavljević, M. Ž. Radojević and A. Miladinović, "Humerus absorbed dose in breast cancer postoperative radiotherapy. Simulation with FOTELP-VOX code and comparison with treatment planning system," *The European Physical Journal Special Topics*, vol. 232, no. 10, p. 1549–1553, 2023.
- [17] M. Zivkovic, D. Miladinović, A. Miladinović, U. Molnar and D. Krstic, "Absorbed dose distribution in human eye simulated by FOTELP-VOX code and verified by volumetric modulated arc therapy treatment plan," 2022.
- [18] J. P. Shah, S. D. Mann, R. L. McKinley and M. P. Tornai, "Characterization of CT Hounsfield units for 3D acquisition trajectories on a dedicated breast CT system," *Journal of X-ray Science and Technology*, vol. 26, no. 4, p. 535–551, 2018.



REVIEW OF METAHEURISTIC APPROACHES FOR BOOSTING MODEL OPTIMIZATION

Miloš Grubješić*,
[0009-0007-3346-8442]

Miodrag Živković,
[0000-0002-4351-068X]

Tamara Živković,
[0000-0003-2969-1709]

Nebojša Bačanin Džakula
[0000-0002-2062-924X]

Singidunum University,
Belgrade, Serbia

Abstract:

Boosting algorithms, recognized for their ability to create strong learners from weak learners, are valuable assets in many areas. Often it is difficult to optimize their hyperparameters, especially when working with high-dimensional datasets. Metaheuristic algorithms, based on natural and evolutionary processes, provide great alternatives to traditional optimization approaches to hyperparameter tuning for boosting models with efficient and robust solutions. This review paper will present an overview of the existing state-of-the-art metaheuristic algorithms utilized in boosting model optimization, highlighting a wide variety of metaheuristic algorithms.

Keywords:

Boosting Algorithms, Metaheuristic Algorithms, Hyperparameter Optimization.

INTRODUCTION

Metaheuristics are a class of stochastic algorithms generated by nature, notably the social behavior of animals. They are successfully deployed to solve NP-hard problems [1]. Metaheuristics provide sufficient outcomes in scenarios where deterministic methods cannot be adopted due to resource or time restrictions, large datasets, nonlinear and/or unstructured problems.

Boosting methods such as XGBoost, LightGBM, CatBoost, and others are capable ranked ML algorithms that are built sequentially upon weak learners [2]. In the end, they create a highly accurate model with defined error minimizations.

Synergizing metaheuristics with boosting methods can vastly improve the model's ability to optimize machine learning models where large complex search space will be searched efficiently, contributing to increased accuracy and robustness of these models.

This paper aims to present a detailed review of metaheuristic approaches merged with boosting methods that showcase their ability to optimize machine learning models regardless of discipline. This review will synthesize takeaways from the critical consideration of selected studies with a metaheuristics-boosting optimization consideration

Correspondence:

Miloš Grubješić

e-mail:

milos.grubjesic.24@singimail.rs



and provide a review of the consideration method, problems, disciplines, and results realized to contribute to enhancing understanding and ways of addressing existing issues in research.

2. OVERVIEW OF BOOSTING AND METAHEURISTICS

2.1. BOOSTING ALGORITHMS

Boosting is an ensemble learning strategy that has become an important part of modern machine learning, combining multiple weak learners to create a strong learner with high predictive accuracy. It operates iteratively, with each successive model attempting to correct the mistakes of the models that preceded it, thereby improving the ensemble's overall predictive performance iteratively. This iterative process allows boosting algorithms to achieve higher accuracy than individual models. Boosting methods are particularly valuable in modeling datasets that exhibit large, complex, and non-linear relationships.

2.2. POPULAR BOOSTING ALGORITHMS

Several boosting algorithms have gained popularity due to their effectiveness and versatility:

- **Adaptive Boosting:** AdaBoost is one of the most widely used boosting algorithms. It assigns weights to each data point in the training set based on the accuracy of prior models and then trains a new model using the updated weights.
- **Gradient Boosting Machines:** GBM is a generalization of AdaBoost. It optimizes.
- **XGBoost:** eXtreme Gradient Boosting is an optimized gradient boosting algorithm known for its efficiency and performance. It includes algorithmic enhancements that contribute to scalability and higher accuracy levels.
- **LightGBM:** Light Gradient Boosting Machine is a gradient boosting framework designed for high performance and efficiency.
- **CatBoost:** CatBoost is a gradient boosting algorithm that excels in handling categorical features. It uses a special technique to deal with categorical variables, often leading to better performance, especially when dealing with such data

2.3. HYPERPARAMETER TUNING

There are various types of hyperparameters that can be tuned to achieve optimal model performance. Some include:

- **Structural parameters:** These control the complexity of individual weak learners, such as the maximum depth of a tree.
- **Learning rate parameters:** These control how much each weak learner contributes to the final model, such as a learning rate.
- **Regularization parameters:** These penalize complex models to limit overfitting. Examples include gamma, alpha, and lambda.
- **Ensemble parameters:** These control the number of weak learners in the ensemble, such as the number of estimators.

2.4. METAHEURISTIC ALGORITHMS

Metaheuristic algorithms can be broadly defined as optimization schemes inspired by deviations in the natural or physical world. "Deviations," in this case, refer to natural inconsistencies or irregularities, deviations that occur in biological, physical, or other natural systems—the deviations from expected or uniform outcomes. Metaheuristic algorithms often use these deviations (such as, for example, random mutations in genetic algorithms or temperature fluctuations in simulated annealing) to diversify the search for solutions and avoid being trapped in sub-optimal areas of the solution space [3]. Their goal is to find a usable solution to a problem that is complex enough that traditional methods would not apply due to constraints of time or resources. Compared to traditional heuristics specific to problems, metaheuristic algorithms provide a higher level of abstraction and are easily applied for wider-reaching optimization problems.

Some notable metaheuristic algorithms include [4]:

- Genetic Algorithms
- Particle Swarm Optimization
- Ant Colony Optimization
- Simulated Annealing
- Tabu Search
- Firefly Algorithm
- Grey Wolf Optimizer, and many more...



These algorithms strive to balance excursion (diversification) and intensification to search for near-optimal solutions within the search space effectively. The general perspective of the excursion is to produce a diverse solution set to examine potential global search space. At the same time, intensification is generally concerned with exploiting the information around the local optimal solution space. Although there is no guarantee that a metaheuristic algorithm will yield an absolute optimal solution each time, a metaheuristic typically produces a near-optimal solution in a reasonable amount of time.

3. REVIEW OF RECENT LITERATURE

A particular emphasis was placed on analysing newer studies from various authors, ensuring the review captures the latest advancements and methodologies in the field.

In [1], the authors integrated XGBoost and AdaBoost with a modified particle swarm optimization algorithm to classify respiratory conditions using a dataset of 920 labelled recordings from 126 patients. This two-tier framework, combining CNN-based feature extraction with boosting and metaheuristics, achieved 98.14% accuracy for binary classification and 81.25% for multi-class classification.

In [5], the researchers introduced an approach that employs XGBoost together with a Genetically Inspired RSA (GIRSA) algorithm for a real-world IoT MQTT dataset comprising six classes of traffic. The method leverages CNNs for feature reduction and obtains an 87.94% multi-class classification accuracy, further enhanced by a modified reptile search algorithm for hyperparameter optimization.

In [6], the authors combined CatBoost and LightGBM with the Chimp optimization algorithm on the CICI-oT2023 intrusion detection dataset. By merging CNN-based feature extraction with boosting models fine-tuned via metaheuristics, the study achieved a 99.83% accuracy rate in multi-class IoT intrusion detection.

In [7], the researchers paired XGBoost with a Coyote optimization algorithm to classify 14,878 tweets in a Twitter sexist harassment dataset. The improved Coyote Optimization Algorithm (IBCOA) outperformed other advanced metaheuristics in tuning XGBoost for more accurate detection of sexist content.

In [8], a data-driven train delay prediction framework was developed by combining XGBoost with Genetic Algorithm, Particle Swarm Optimization, Whale Optimization Algorithm, and Grey Wolf Optimization. Using data from the Beijing–Shanghai high-speed rail-

way, the model delivered high accuracy, low prediction errors, and enhanced interpretability by analysing dispatching commands and delay propagation mechanisms.

In [9], the authors used LightGBM optimized by a Whale Optimization (WO) algorithm to diagnose thyroid disease, achieving a 99.75% accuracy. This proves the efficacy of combining boosting models and metaheuristics for medical classification tasks.

In [10], researchers explored both XGBoost and LightGBM alongside Genetic Algorithm and Simulated Annealing on the Pima Indian Diabetes dataset. LightGBM with Genetic Algorithm provided the best performance at 86% accuracy, emphasizing the importance of effective feature selection in classification models.

In [11], XGBoost was integrated with a Modified Boxing Match (MBM) algorithm to forecast electricity consumption. The findings show that incorporating MBM significantly improved the model's accuracy for electricity demand prediction.

In [12], the authors combined XGBoost with an Artificial Bee Colony (ABC) algorithm for sales forecasting across three open-source datasets. A new hybrid approach—merging ABC with the Fire Hawk Optimizer (FHO)—yielded superior performance metrics (RMSE and MAPE) compared to using ABC or FHO alone.

In [13], XGBoost was paired with Particle Swarm Optimization and Gray Wolf Optimization to predict failure time for nine shovels at the Gol-Gohar iron ore mine in Iran. The PSO–XGB approach achieved an R^2 of 0.99 and enabled predictive maintenance, saving an estimated \$61,189 per month.

In [14], an Enhanced Gradient Boosting Machine (EGBM) was proposed, optimized with a modified Particle Swarm Optimization, for customer churn prediction on seven open-source datasets. This CP-EGBM model outperformed traditional models like GBM and SVM, delivering high accuracy, recall, F1-measure, and AUC values in the telecommunications sector.

In [15], a hybrid SAOA–LightGBM method was presented to improve fault warning accuracy in industrial settings using 1,500 samples from a supervisory information system. With a 90% fault warning accuracy rate, it surpassed comparable state-of-the-art models in both prediction accuracy and generalization ability.

In [16], XGBoost, enhanced by a Boosted Particle Swarm Optimization algorithm, was applied to publicly available AIS data from Kaggle. The approach attained a 99.72% overall accuracy in vessel classification, illustrating the effectiveness of combining boosting and metaheuristics.



In [17], CatBoost was paired with the Grey Wolf Optimizer (GWO) to assess landslide susceptibility in Kerala, India, using data gathered from DEM, data portals, and published maps. The resulting CatBoost–GWO model showed a high AUC value of 0.910, highlighting robust predictive capability.

In [18], researchers optimized XGBoost using a Repetilian search algorithm to analyse a two-year database of hourly pollutant concentrations, including toluene. The study demonstrated how advanced metaheuristics paired with boosting models can significantly improve environmental data analytics.

In [19], a teaching–learning-based (TLB) optimization technique was implemented alongside XGBoost to detect fraud in a synthetic credit card dataset containing transactions from European cardholders. The hybrid model outperformed existing techniques by effectively tuning the boosting classifier and providing superior predictive performance.

In [20], a Hybrid Adaptive Red Fox Optimization algorithm (HARFO) was implemented alongside XGBoost and AdaBoost to detect insider threats through sentiment classification in the Insider Threat Test Dataset provided by the Carnegie Mellon University Software Engineering Institute. This synthetic dataset simulates logs of a large business over a 500-day interval, and the metaheuristic-enhanced approach produced commendable outcomes across various simulated scenarios by effectively tuning the boosting classifiers for superior predictive performance.

In [21], an iterative sine-cosine metaheuristic algorithm was integrated with an XGBoost model to enhance the prediction of external audit opinions. Leveraging a dataset of 12,690 observations from Serbian companies (2016–2019) and incorporating 598 variables, the approach aimed to surpass previously established benchmarks. By comparing six different metaheuristics for hyperparameter tuning and evaluating two distinct test scenarios with varying levels of difficulty, the optimized XGBoost model demonstrated superior predictive performance.

In [22], a framework based on a two-tier structure was presented, which uses a convolutional neural network (CNN) throughout the first layer and AdaBoost/XGBoost classifiers in the second layer, with PSO (particle swarm optimisation) implemented for hyperparameter optimisation. 5 data sets from NASA's Promise repository (KC1, JM1, CM1, KC2, and PC1) and approximately 1000 crowd-sourced Python programming problems were used. The two-tier structure produced

accuracies of the classifiers of 0.768799 and 0.772166 while NLP solutions demonstrated even higher accuracies of 0.979781 and 0.983893 from the AdaBoost and XGBoost, respectively, thereby surpassing existing solutions.

4. CONCLUSION

These studies highlight the powerful combination of boosting algorithms and metaheuristic optimization for various tasks. Whether applied to IoT security, medical diagnosis, infrastructure maintenance, or sales and demand forecasting, this combined approach consistently demonstrates notable improvements in predictive performance.

Metaheuristic algorithms (e.g., modified reptile search, improved Coyote Optimization Algorithm, hybrid ABC-FHO, MBM, PSO, GWO) have been shown to be instrumental in fine-tuning the hyperparameters of boosting models, offering clear gains over traditional or standalone optimization methods.

Although the studies provided evidence of strengthening boosting by metaheuristic optimization, there are still areas to further investigate. For instance, the effectiveness of approaches may depend on the context in which the approach is utilized or the nature of the datasets that were used. By conducting tests on the same dataset and establishing benchmarks, one could determine whether the improvements were consistent across the varying contexts, allowing for a clearer identification of the best-performing approaches.

In general, the evidence clearly indicates that integrating metaheuristic-based hyperparameter tuning with advanced boosting algorithms generates robust, high-performing models that can adapt to a range of complex real-world problems.



REFERENCES

- [1] S. Purkovic et al., "Audio analysis with convolutional neural networks and boosting algorithms tuned by metaheuristics for respiratory condition classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 10, p. 102261, Dec. 2024, doi: 10.1016/j.jksuci.2024.102261.
- [2] 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, in KDD '16*. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [3] I. Boussaïd, J. Lepagnot, and P. Siarry, "A survey on optimization metaheuristics," *Elsevier BV*, vol. 237, pp. 82–117, Mar. 2013, doi: 10.1016/j.ins.2013.02.041.
- [4] S. Almufti, R. Ali, and J. Fuente, "Overview of Metaheuristic Algorithms," *Polaris Global Journal of Scholarly Research and Trends*, vol. 2, pp. 10–32, Apr. 2023, doi: 10.58429/pgjsrt.v2n2a144.
- [5] M. Salb et al., "Enhancing Internet of Things Network Security Using Hybrid CNN and XGBoost Model Tuned via Modified Reptile Search Algorithm," *Applied Sciences*, vol. 13, no. 23, p. 12687, Jan. 2023, doi: 10.3390/app132312687.
- [6] M. Antonijevic et al., "Intrusion detection in metaverse environment internet of things systems by metaheuristics tuned two level framework," *Sci Rep*, vol. 15, no. 1, p. 3555, Jan. 2025, doi: 10.1038/s41598-025-88135-9.
- [7] M. Dobrojevic et al., "Cyberbullying Sexism Harassment Identification by Metaheuristics-Tuned eXtreme Gradient Boosting," *EBSCOhost*, Sep. 2024. doi: 10.32604/cmc.2024.054459.
- [8] T. Gao, J. Chen, and H. Xu, "Data-driven train delay prediction incorporating dispatching commands: An XGBoost-metaheuristic framework," *IET Intelligent Transport Systems*, vol. 18, no. 10, pp. 1777–1796, 2024, doi: 10.1049/itr2.12461.
- [9] B. B. Sinha, M. Ahsan, and R. Dhanalakshmi, "LightGBM empowered by whale optimization for thyroid disease detection," *International Journal of Information Technology*, vol. 15, no. 4, pp. 2053–2062, Apr. 2023, doi: 10.1007/s41870-023-01261-3.
- [10] G. Alp and F. Soygazi, "Meta-Heuristic Supported Feature Selection in Classification Algorithms for Diabetes Diagnosis," in *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, Mersin, Türkiye: IEEE, May 2024, pp. 1–4. doi: 10.1109/SIU61531.2024.10601062.
- [11] N. Liao, Z. Hu, and D. Magami, "A metaheuristic approach to model the effect of temperature on urban electricity need utilizing XGBoost and modified boxing match algorithm," *AIP Adv*, vol. 14, no. 11, p. 115318, Nov. 2024, doi: 10.1063/5.0219713.
- [12] B. Gülsün, "Optimizing a Machine Learning Algorithm by Novel Metaheuristic Approach: Case Study in Forecasting," Nov. 13, 2024, Preprints: 2024110907. doi: 10.20944/preprints202411.0907.v1.
- [13] M. Mirzehi Kalateh Kazemi, Z. Nabavi, M. Rezakhah, and A. Masoudi, "Application of XGB-based metaheuristic techniques for prediction time-to-failure of mining machinery," *Systems and Soft Computing*, vol. 5, p. 200061, Dec. 2023, doi: 10.1016/j.sasc.2023.200061.
- [14] I. AlShourbaji, N. Helian, Y. Sun, A. G. Hussien, L. Abualigah, and B. Elnam, "An efficient churn prediction model using gradient boosting machine and metaheuristic optimization," *Sci Rep*, vol. 13, no. 1, p. 14441, Sep. 2023, doi: 10.1038/s41598-023-41093-6.
- [15] S. Li, N. Jin, A. Dogani, Y. Yang, M. Zhang, and X. Gu, "Enhancing LightGBM for Industrial Fault Warning: An Innovative Hybrid Algorithm," *Processes*, vol. 12, no. 1, p. 221, Jan. 2024, doi: 10.3390/pr12010221.
- [16] A. Petrovic et al., "Marine Vessel Classification and Multivariate Trajectories Forecasting Using Metaheuristics-Optimized eXtreme Gradient Boosting and Recurrent Neural Networks," *Applied Sciences*, vol. 13, no. 16, p. 9181, Jan. 2023, doi: 10.3390/app13169181.
- [17] R. S. Ajin, S. Segoni, and R. Fanti, "Optimization of SVR and CatBoost models using metaheuristic algorithms to assess landslide susceptibility," *Sci Rep*, vol. 14, no. 1, p. 24851, Oct. 2024, doi: 10.1038/s41598-024-72663-x.
- [18] A. Stojić, "Revealing Toluene Behaviour in the Atmosphere Based on Coupling of Metaheuristics, Xgboost, and Shap," *Sinteza 2023 - International Scientific Conference on Information Technology, Computer Science, and Data Science*, pp. 17–22, 2023, doi: 10.15308/Sinteza-2023-17-22.
- [19] A. Petrovic, M. Antonijevic, I. Strumberger, L. Jovanovic, N. Savanovic, and S. Janicijevic, "The XGBoost Approach Tuned by TLB Metaheuristics for Fraud Detection," in *Proceedings of the 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022)*, N. Bacanin and H. Shaker, Eds., Dordrecht: Atlantis Press International BV, 2023, pp. 219–234. doi: 10.2991/978-94-6463-110-4_16.



- [20] D. Mladenovic et al., "Sentiment classification for insider threat identification using metaheuristic optimized machine learning classifiers," *Sci Rep*, vol. 14, no. 1, p. 25731, Oct. 2024, doi: 10.1038/s41598-024-77240-w.
- [21] M. Todorovic, N. Stanisic, M. Zivkovic, N. Bacanin, V. Simic, and E. B. Tirkolaee, "Improving audit opinion prediction accuracy using metaheuristics-tuned XGBoost algorithm with interpretable results through SHAP value analysis," *Applied Soft Computing*, vol. 149, p. 110955, Dec. 2023, doi: 10.1016/j.asoc.2023.110955.
- [22] A. Petrovic et al., "Exploring Metaheuristic Optimized Machine Learning for Software Defect Detection on Natural Language and Classical Datasets," *Mathematics*, vol. 12, no. 18, p. 2918, Sep. 2024, doi: 10.3390/math12182918.



INTRUSION DETECTION TECHNIQUES AND SWARM INTELLIGENCE CYBERSECURITY REVIEW

Zorana Krsmanović*,
[0009-0005-5642-8190]

Stojanka Tešanović,
[0009-0005-2371-3358]

Aleksandar Petrović,
[0000-0003-3324-3909]

Miodrag Živković,
[0000-0002-4351-068X]

Tamara Živković
[0000-0003-2969-1709]

Singidunum University,
Belgrade, Serbia

Abstract:

A computing and communications revolution at high speed has hastened the demand for effective security devices to protect networks from highly sophisticated cyberattacks. Intrusion Detection Systems (IDS) are an essential part of network traffic monitoring and network abuse detection. Traditional IDS techniques, however, such as signature-based and anomaly-based systems, experience severe limitations, including weak detection of novel attacks, high false positives, and high computational overhead. This survey provides a comprehensive overview of state-of-the-art hybrid machine learning (ML) methods with swarm intelligence (SI), a collection of metaheuristic optimization techniques inspired by collective behaviour in nature, for the enhancement of IDS. The examination is critical and covers hybrid models integrating supervised, unsupervised, and deep learning methods optimized using SI methods, such as crayfish optimization, firefly algorithm (FA), and social network search (SNS). Their key strengths and weaknesses and their applications in the real world are highlighted. Problems of computational complexity, scalability, and real-time use are also cited. The paper identifies critical areas for future research activity, such as improved feature selection methodology, real-time adaptability, distributed processing methodology, and large and diverse benchmark datasets. The survey highlights the immense scope for hybrid SI-based ML solutions to improve cybersecurity practice and research.

Keywords:

Intrusion Detection, Swarm Intelligence, Machine Learning, Metaheuristics, Cybersecurity.

INTRODUCTION

In today's era of rapid digital systems development, malicious activities constantly attempt to compromise data integrity, driven by financial gain, data theft, or other malicious motives. For this reason, special emphasis is placed on securing data and system infrastructure, where intrusion detection methods are used to identify unauthorized access. However, as these methods become more advanced, attackers continuously develop new techniques that must be countered to protect sensitive data from compromise. Intrusion detection can be classified into two categories, signature-based and anomaly-based techniques. Signature-based intrusion detection techniques rely on predefined rules and known attack patterns from an existing database but cannot detect unknown threats,

Correspondence:

Zorana Krsmanović

e-mail:

zorana.krsmanovic23@singimail.rs



while anomaly-based techniques monitor network activity for deviations using methods like clustering and classification, both facing challenges from increasing data volume, and malicious behaviour, and the need to process numerous attributes [1]. Such traditional systems rely on previously known attacks and cannot effectively detect unknown, new, and sophisticated threats.

Although various methods based on statistics and machine learning (ML) exist, many still face challenges such as high false alarm rates and low accuracy in detecting new attacks [2]. Additionally, slow convergence and the lack of efficient techniques for optimization and feature selection are also challenges for intrusion detection [3]. Furthermore, many require significant computational resources and fail to find a balance between accuracy and efficiency, making them impractical for real-world implementation [4]. These challenges highlight the need for developing more advanced methods capable of better detecting threats in modern computer networks.

The applications of artificial intelligence (AI) and ML techniques are increasingly being used in the development of intrusion detection systems (IDS). According to Heidari et al. [5], ML is becoming a key component in network intrusion detection systems (NIDS) and intrusion prevention systems (IPS), providing greater accuracy compared to traditional rule-based methods. This shift is the result of the development of hardware accelerators and sophisticated ML algorithms [6], [7], [8], enabling more precise detection of network breaches and more efficient analysis of network traffic. AI-based systems have shown exceptional performance in anomaly detection and threat classification. The main goal initially was the implementation of traditional ML models such as decision tree (DT) [9] and support vector machine (SVM) in intrusion detection systems, to later introduce deep learning methods such as convolutional neural networks (CNN), long short-term memory networks (LSTM), and autoencoders [10]. However, despite significant potential, the application of such solutions in real operational environments still presents numerous challenges [11]. Therefore, a detailed review and analysis of state-of-the-art hybrid ML approaches for intrusion detection are essential for better understanding the real capabilities and limitations of AI.

Swarm intelligence (SI) is a branch of AI that focuses on optimization using metaheuristic methods inspired by collective behaviour in nature. Algorithms from the SI group rely on the idea that a group of individuals can find a better solution than a single individual. Each indi-

vidual or agent in an SI algorithm represents a potential solution. Through constant interaction and exchange of information with other agents, the group evolves towards increasingly better solutions. This principle is inspired by the behaviour of animal communities, such as ants that communicate using pheromones while searching for food, marking the shortest and most efficient paths, or birds that move in flocks, adjusting their trajectory based on the position of their neighbours [12]. Thus, the basic principle is that many potential solutions are considered simultaneously and adjusted during each iteration, with collective behaviour finding a suboptimal global solution. The optimal solution is considered to be the best possible and often it is very hard and even impossible to reach. Hence approximation techniques target the suboptimal solution, which is considered to be very close to the optimal but it is reachable in polynomial time. Swarm Intelligence algorithms are based on two phases: exploration and exploitation. In the exploration phase, the algorithm randomly searches the entire search space, i.e., the space of all possible solutions. Each individual moves randomly through the space. The goal here is to ensure that the algorithm does not get stuck in the local optimum but explores a wide range of solutions. On the other hand, exploitation focuses on the most promising region of the search space and deepens the search within it. A balance between these two phases must exist so that the algorithm does not remain trapped in a local optimum or aimlessly wander through the entire search space [13]. One way to achieve a better balance is the process of hybridization, where the global search capability of one algorithm is combined with the efficient local search of another [14]. This approach is in line with the no-free lunch theorem (NFL) [15], which implies that no single optimization algorithm can outperform all others in every scenario. Because of this, combining different strategies is crucial for achieving better overall performance.

2. RELATED WORKS

As cyber threats are becoming more sophisticated, IDS have struggled due to their reliance on fixed signatures, limited detection of unknown attacks, and tendency toward high false positives. As a result, researchers are now exploring hybrid and adaptive approaches to make IDS more flexible and effective in real-time environments. Meanwhile, SI, inspired by natural group behaviours such as those of insects and birds, is emerging as a promising method for solving complex



optimization problems. These algorithms effectively balance broad exploration and focused searching, showing great potential in fields like cybersecurity, wireless sensor networks, medical diagnostics, and cryptocurrency forecasting.

2.1. INTRUSION DETECTION SYSTEMS

In today's rapidly evolving cyber landscape, it's clear that conventional network security methods struggle to keep pace with increasingly sophisticated and frequent attacks. Therefore, IDS, as a security tool, addresses these limitations by monitoring both internal and external network activities [16], [17]. They are used to observe network traffic for harmful actions, including data theft, censorship, or violation of network protocols. Traditional IDS solutions include network-based (NIDS), host-based (HIDS), wireless (WIDS), and network behaviour analysis (NBA) systems, employing signature-based (SIDS), anomaly-based (AIDS), or stateful protocol analysis (SPA) methodologies for threat detection.

Conventional approaches have demonstrated their effectiveness in detecting particular kinds of cyber threats, especially when utilizing established attack patterns. Traditional signature-based IDS effectively detect known cyber threats by comparing incoming packets with predefined signatures, but Kumar et al. [18] highlight their ineffectiveness against unknown or modified threats due to dependency on existing signature databases. Although effective in recognizing novel threats by identifying deviations from normal network activity, AIDS generates higher false-positive rates, increasing workload for security analysts, as noted by Narsingyani et al. [19].

Detection systems relying on protocol analysis, such as SPA, identify deviations from standard protocol behaviours, providing strong protection against unauthorized protocol usage and attacks. According to Nitin et al. [20], SPA is resource-intensive and vulnerable to advanced threats designed to mimic legitimate protocol behaviours. While NBA statistically profiles network traffic to detect significant threats like DDoS attacks, Moon et al. [21] point out its limitations in identifying subtle or covert cyber threats. As wireless communication becomes more widely used, WIDS has emerged as a critical part of network security, providing targeted monitoring for wireless traffic and detecting threats that are specific to wireless networks. Afzal et al. [22] emphasize that WIDS face difficulties in detecting passive monitoring attacks because of their dependence on static signatures, underlining the need for adaptive security approaches.

Hybrid IDS solutions, merging signature-based accuracy with anomaly-based flexibility, showcase the optimal balance of precision and resource effectiveness. Nonetheless, they demand considerable computational power, rendering them less appropriate for real-time use in resource-limited settings [23]. These findings emphasize the importance of enhancing IDS methods by incorporating explainability and efficiency to address the limitations of conventional strategies.

Even though they remain significant, traditional IDS systems encounter increasing difficulties in today's cybersecurity environments. One significant disadvantage is their computational and hardware constraints, especially for signature-based IDS, which necessitate regular database updates to stay effective against new threats like zero-day attacks [24]. This dependence on predetermined attack patterns limits their flexibility, rendering them susceptible to fast-changing malware. Anomaly-based IDS is more versatile yet faces scalability challenges because of its elevated false positive rate, requiring significant adjustments for use in dynamic network settings [25].

2.2. SWARM INTELLIGENCE

Metaheuristic techniques inspired by collective behaviours found in nature, particularly those observed in animals exhibiting swarm-like activities, form the basis of SI algorithms. Due to their effectiveness in addressing NP-hard problems, these algorithms have become highly popular for optimization purposes. Their performance can be further enhanced through the process of hybridization. This procedure involves the integration of various algorithms to harmonize their unique strengths and limitations. This hybridization is particularly valuable because swarm intelligence algorithms typically excel either in exploration (broadening the search space) or in exploitation (focusing on precise solutions), and finding the right balance between these two phases is critical for achieving optimal results [26], [27]. Despite their extensive applicability, the NFL theorem posits that no singular algorithm is able to consistently outperform all others. Applications cover areas including wireless sensor networks [28-29], cryptocurrency forecasting [30], optimization of neural networks [31], cloud-edge computing [32], and medical diagnostics [33].



3. APPLICATION OF ARTIFICIAL INTELLIGENCE IN CYBERSECURITY

The impact of AI on cybersecurity is profound and far-reaching. The use of AI technologies enables rapid detection of malicious behaviours for security teams to respond quickly before any real damage occurs. Applications powered by AI-driven technology can search through enormous amounts of data in real time, improving the detection of anomalies and possible threats [34]. Besides detection, AI is also used to predict future cyberattacks using behaviour patterns so that proactive security can be implemented. The development of artificial neural networks has improved malware analysis and made cloud security far more robust. However, AI in cybersecurity also presents challenges, as ethical concerns and the possibility of cybercriminals utilizing AI for negative purposes are significant threats [35].

The application of metaheuristic optimization techniques in cybersecurity has significantly enhanced the efficiency of ML models in detecting cyberattacks. Jokić et al. [36] present the application of a crayfish optimization algorithm (COA) and genetic algorithm (GA) for the optimization of extreme gradient boosting (XGBoost) in detecting structured query language (SQL) injection attacks. The method applies natural language processing (NLP) techniques to improve security against SQL-based cyberattacks, highlighting the power of metaheuristics in AI model optimization.

Similarly, Baćanin et al. [37] introduced a modified SNS algorithm for optimizing XGBoost for intrusion detection, emphasizing the advantages of swarm

intelligence approaches in handling NP-hard security problems. Their study highlights how SNS-enhanced XGBoost outperforms conventional tuning techniques in detecting anomalies within internet of things (IoT) networks.

In another study, Savanović et al. [38] apply a modified firefly algorithm (FA) to optimize ML models for intrusion detection in Healthcare 4.0 Internet of Things (IoT) systems. The results indicate that FA-based metaheuristic tuning significantly improves accuracy and detection speed, making it a suitable solution for real-time cyber threat mitigation.

Finally, Živković et al. [39] propose a hybridized sine cosine algorithm (SCA) to fine-tune XGBoost in identifying vulnerabilities in IoT healthcare security, showcasing how metaheuristic approaches can enhance predictive performance and system resilience. These studies collectively underscore the growing importance of metaheuristic optimization in cybersecurity, particularly in hybridized AI-driven security solutions, where advanced optimization techniques ensure both efficiency and adaptability in modern cybersecurity defence frameworks.

4. COMPARATIVE ANALYSIS OF INTRUSION DETECTION TECHNIQUES

A structured comparison of these IDS methods, highlighting their strengths and limitations, is provided in Table 1. This comparison aims to illustrate the trade-offs between detection efficiency, resource consumption, and adaptability to emerging threats.

Table 1. Advantages and disadvantages of different IDS technologies

IDS Method	Reference Citation	Advantages	Disadvantages	Zero-Day Attack Detection	Resource Requirements	False Alarm Rate
Signature-Based IDS (SIDS)	[24]	Fast detection of known threats easy implementation	Cannot detect new threats, needs frequent updates	No	Low	Low
Anomaly-Based IDS (AIDS)	[25]	Can detect unknown threats, learns from behaviour	High false positive rate, difficult fine-tuning	Yes	High	High
Stateful Protocol Analysis	[24]	Effective against protocol deviations	High resource consumption, complex setup	Partially	High	Low
Network Behavior Analysis	[25]	Good for detecting DDoS, a traffic-based approach	Poor detection of low-intensity attacks	Partially	High	Medium
Wireless IDS (WIDS)	[23]	Specialized for wireless threats, enhances Wi-Fi security	Cannot detect passive attacks, limited scope	No	Low	Medium



As shown in Table 1, conventional IDS techniques provide different compromises regarding detection efficacy and computational performance. Signature-based IDS is the most efficient in terms of computation, yet it fails to defend against zero-day attacks because it depends on known signatures [24]. Conversely, anomaly-based IDS offers flexibility, but its elevated false-positive rate may cause alert fatigue in security analysts, diminishing overall efficiency in extensive implementations [25].

An overview of different metaheuristic technologies from the SI group of algorithms that are applied to the field of cybersecurity is provided. These types of algorithms have proven excellent optimizers, and each of the reviewed papers provides a high-performance solution for increasing cybersecurity. The role of SI in cybersecurity is therefore confirmed, and it can only increase as the future of cybersecurity will be AI. In the work of Jokic et al. [36] the authors explore security improvements through SI optimization by employing an SQL injection detection mechanism. Another approach based on the XGBoost algorithm and optimized by the SI algorithm is explored in the work of Bacanin et al. [37]. A different aspect of the intrusion detection is explored in the work of Savanović et al. [38], as the authors tackle the problem with a similar approach but for IoT systems. Lastly, a work by Zivkovic et al. [39] is reviewed in which the authors once more employ SI-optimized intrusion detection in IoT healthcare systems but this time a metaheuristic-math-based optimizer is employed for optimizing XGBoost.

These obstacles underscore the need for ongoing innovation in IDS development, especially in enhancing detection algorithms, minimizing false positives, and increasing computational efficiency. Future research pathways ought to emphasize hybrid and adaptive strategies that reconcile security efficacy with practical performance limitations.

5. CONCLUSION

Systems for detecting and preventing potential threats, known as IDS, play a crucial role in ensuring cybersecurity. This paper examined hybrid ML techniques for intrusion detection, highlighting their advantages over single techniques and traditional approaches. By integrating several techniques, hybrid models provide enhanced detection, fewer false positives, and greater responsiveness to evolving cyber threats.

Effective detection and prevention of cybersecurity attacks are crucial to maintaining network security. In this paper, focus was given to hybrid ML techniques for intrusion detection, pointing out the advantage of these over traditional methods. The study explored various intrusion detection methodologies, citing the limitation of signature-based and anomaly-based methods, especially their inability to detect sophisticated cyberattacks.

A review of recent research confirmed that hybrid solutions, which combine different ML methodologies, offer improved detection rates, fewer false alarms, and flexibility. In addition, the combination of SI and other nature-inspired optimization strategies was experimented and confirmed to improve IDS performance.

Despite advancements, challenges such as computational complexity, real-time processing constraints, and smaller dataset sizes remain. Optimizing hybrid systems through advanced feature selection, real-time tunability, and scalable computation methods would work in the future. Large-scale benchmark datasets, along with an exploration into transfer learning, could improve these systems' generalizability.

It is important to note that the evolving AI regulations use pose threat from cybersecurity as well as from the perspective of general security. Cybersecurity is a critical field even without implicating AI, which only increases the risk of mishap. However, this should not hinder the development of AI and cybersecurity, hence it is important to emphasise that this risk can be mitigated with proper regulations.

Lastly, hybrid ML techniques are a possible prospect for intrusion detection with greater accuracy and dependability. Continued innovation in the field will be essential to further strengthening cybersecurity defences against the ever-emerging world of cyber threats.



REFERENCES

- [1] S. Sharma, V. Kumar, and K. Dutta, "Multi-objective optimization algorithms for intrusion detection in IoT networks: A systematic review," *Internet of Things and Cyber-Physical Systems*, 2024. doi: 10.1016/j.iotcps.2024.01.003
- [2] Y. Al Sawafi, A. Touzene, and R. Hedjam, "Hybrid deep learning-based intrusion detection system," unpublished, 2023. doi: 10.2139/ssrn.3994183
- [3] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowl.-Based Syst.*, vol. 189, p. 105124, 2020. doi: 10.1016/j.knsys.2019.105124
- [4] M. H. Nasir, S. A. Khan, M. M. Khan, and M. Fatima, "Swarm intelligence inspired intrusion detection systems—A systematic literature review," *Comput. Netw.*, vol. 205, p. 108708, 2022. doi: 10.1016/j.comnet.2021.108708
- [5] A. Heidari, N. Navimipour, M. Unal, and G. Zhang, "Machine learning applications in Internet-of-Drones: Systematic review, recent deployments, and open issues," *ACM Comput. Surv.*, vol. 55, 2022, doi: 10.1145/3571728.
- [6] S. Bukhari et al., "Secure and privacy-preserving intrusion detection in wireless sensor networks: Federated learning with SCNN-Bi-LSTM for enhanced reliability," *Ad Hoc Netw.*, vol. 155, p. 103407, 2024, doi: 10.1016/j.adhoc.2024.103407.
- [7] A. Hanafi et al., "Intrusion detection in Internet of Things using improved binary golden jackal optimization algorithm and LSTM," *Cluster Comput.*, vol. 27, pp. 1–18, 2023, doi: 10.1007/s10586-023-04102-x.
- [8] M. Belouch and S. El Hadaj, "Comparison of ensemble learning methods applied to network intrusion detection," in *Proc. ACM Int. Conf. Computer Systems and Technologies*, 2017, pp. 1–4, doi: 10.1145/3018896.3065830.
- [9] J. R. Quinlan, C4.5: Programs for Machine Learning, San Mateo, CA, USA: Morgan Kaufmann, 2014.
- [10] M. Sajid et al., "Enhancing intrusion detection: A hybrid machine and deep learning approach," *J. Cloud Comput.*, vol. 13, no. 1, p. 123, 2024. doi: 10.1186/s13677-024-00685-x
- [11] I. Goodfellow, Deep Learning, Cambridge, MA, USA: MIT Press, 2016.
- [12] C. Wang et al., "Swarm intelligence: A survey of model classification and applications," *Chin. J. Aeronaut.*, 2024, Art. no. 102982. doi: 10.1016/j.cja.2024.03.019
- [13] N. Bacanin et al., "Performance of a novel chaotic firefly algorithm with enhanced exploration for tackling global optimization problems: Application for dropout regularization," *Mathematics*, vol. 9, p. 2705, 2021, doi: 10.3390/math9212705.
- [14] R. Shankar et al., "Hybridized particle swarm–gravitational search algorithm for process optimization," *Processes*, vol. 10, no. 3, p. 616, 2022. doi: 10.3390/pr10030616
- [15] T. F. Sterkenburg and P. D. Grünwald, "The no-free-lunch theorems of supervised learning," *Synthese*, vol. 199, no. 3, pp. 9979–10015, 2021. doi: 10.1007/s11229-021-03233-1
- [16] K. Scarfone and P. Mell, Guide to Intrusion Detection and Prevention Systems (IDPS), NIST Special Publication 800-90, 2007.
- [17] F. Sabahi and A. Movaghar, "Intrusion detection: A survey," in *Proc. 3rd Int. Conf. Syst. Netw. Commun.*, Oct. 2008, pp. 23–26. doi: 10.1109/ICSNC.2008.44
- [18] V. Kumar and O. P. Sangwan, "Signature based intrusion detection system using SNORT," *Int. J. Comput. Appl. Inf. Technol.*, vol. 1, no. 3, pp. 35–41, Nov. 2012.
- [19] D. Narsingyani and O. Kale, "Optimizing false positive in anomaly based intrusion detection using genetic algorithm," in *Proc. IEEE 3rd Int. Conf. MOOCs, Innov. Technol. Educ. (MITE)*, Oct. 2015, pp. 72–77. doi: 10.1109/MITE.2015.7375291
- [20] T. Nitin, S. R. Singh, and P. G. Singh, "Intrusion detection and prevention system (IDPS) technology-network behavior analysis system (NBAS)," *ISCA J. Eng. Sci.*, vol. 1, no. 1, pp. 51–56, 2012.
- [21] D. Moon, H. Im, I. Kim, and J. H. Park, "DTB-IDS: An intrusion detection system based on decision tree using behavior analysis for preventing APT attacks," *J. Supercomput.*, vol. 73, no. 7, pp. 2881–2895, 2017. doi: 10.1007/S11227-015-1604-8
- [22] Z. Afzal, J. Rossebø, B. Talha, and M. Chowdhury, "A wireless intrusion detection system for 802.2 networks," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2016, pp. 828–834.
- [23] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, "Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 112392–112415, 2022. doi: 10.1109/WiSPNET.2016.7566249
- [24] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, 2019. doi: 10.1186/s42400-019-0038-7
- [25] A. Khraisat and A. Alazab, "A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges," *Cybersecurity*, vol. 4, pp. 1–27, 2021. doi: 10.1186/s42400-021-00077-7



- [26] R. Poli, J. Kennedy, and T. Blackwell, "Swarm intelligence," *Particle Swarm Optim.*, vol. 1, no. 1, pp. 33–57, 2007.
- [27] D. Karaboga, "Artificial bee colony algorithm," *Scholarpedia*, vol. 5, no. 3, p. 6915, 2010. doi: 10.4249/scholarpedia.6915
- [28] N. Bacanin, E. Tuba, M. Zivkovic, I. Strumberger, and M. Tuba, "Whale optimization algorithm with exploratory move for wireless sensor networks localization," in *Proc. Int. Conf. Hybrid Intelligent Syst.*, Cham, Switzerland: Springer, Dec. 2019, pp. 328–338. doi: 10.1007/978-3-030-49336-3_33
- [29] M. Zivkovic, N. Bacanin, E. Tuba, I. Strumberger, T. Bezdan, and M. Tuba, "Wireless sensor networks life time optimization based on the improved firefly algorithm," in *Proc. 2020 Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2020, pp. 1176–1181. doi: 10.1109/IWCMC48107.2020.9148087
- [30] M. Salb, M. Zivkovic, N. Bacanin, A. Chhabra, and M. Suresh, "Support vector machine performance improvements for cryptocurrency value forecasting by enhanced sine cosine algorithm," in *Proc. CVR 2021*, Singapore: Springer, 2022, pp. 527–536. doi: 10.1007/978-981-16-8225-4_40
- [31] N. Bacanin, T. Bezdan, M. Zivkovic, and A. Chhabra, "Weight optimization in artificial neural network training by improved monarch butterfly algorithm," in *Proc. ICMCSI 2021*, Singapore: Springer, 2022, pp. 397–409. doi: 10.1007/978-981-16-1866-6_29
- [32] N. Bacanin, M. Zivkovic, T. Bezdan, K. Venkatachalam, and M. Abouhawwash, "Modified firefly algorithm for workflow scheduling in cloud-edge environment," *Neural Comput. Appl.*, vol. 34, no. 11, pp. 9043–9068, 2022. doi: 10.1007/s00521-022-06925-y
- [33] T. Bezdan, M. Zivkovic, E. Tuba, I. Strumberger, N. Bacanin, and M. Tuba, "Glioma brain tumor grade classification from MRI using convolutional neural networks designed by modified FA," in *Proc. Int. Conf. Intell. Fuzzy Syst.*, Cham, Switzerland: Springer, Jul. 2020, pp. 955–963. doi: 10.1007/978-3-030-51156-2_111
- [34] A. Oseni, N. Moustafa, H. Janicke, P. Liu, Z. Tari, and A. Vasilakos, "Security and privacy for artificial intelligence: Opportunities and challenges," *arXiv preprint arXiv:2102.04661*, 2021. doi: 10.48550/arXiv.2102.04661
- [35] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Pratharaj, "From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy," *IEEE Access*, vol. 11, pp. 80218–80245, 2023. doi: 10.1109/ACCESS.2023.3300381
- [36] A. Jokic, N. Jovic, V. Gajic, and M. S. Milos, "Structured query language injection detection with natural language processing techniques optimized by metaheuristics," in *Proc. 2nd Int. Conf. Innovation Inf. Technol. Bus. (ICIITB 2024)*, Springer Nature, vol. 113, p. 155, Aug. 2024. doi: 10.2991/978-94-6463-482-2_11
- [37] N. Bacanin, A. Petrovic, M. Antonijevic, M. Zivkovic, M. Sarac, E. Tuba, and I. Strumberger, "Intrusion detection by XGBoost model tuned by improved social network search algorithm," in *Proc. Int. Conf. Modelling Dev. Intell. Syst.*, Cham, Switzerland: Springer Nature, Oct. 2022, pp. 104–121. doi: 10.1007/978-3-031-27034-5_7
- [38] N. Savanović, A. Toskovic, A. Petrovic, M. Zivkovic, R. Damaševičius, L. Jovanovic, et al., "Intrusion detection in Healthcare 4.0 Internet of Things systems via metaheuristics optimized machine learning," *Sustainability*, vol. 15, p. 12563, 2023. doi: 10.3390/su151612563
- [39] M. Zivkovic, L. Jovanovic, N. Bacanin, A. Petrovic, N. Savanovic, and M. Dobrojevic, "XGBoost tuned by hybridized SCA metaheuristics for intrusion detection in Healthcare 4.0 IoT systems," in *Proc. Int. Conf. Eng., Appl. Sci. Syst. Model.*, Singapore: Springer, Apr. 2017, pp. 1–16. doi: 10.1007/978-981-99-8438-1_1



DATA SCIENCE MEETS FRAGRANCE: ANALYZING USER REVIEWS TO DECODE EMOTIONAL CONNECTIONS TO PERFUME NOTES

Milena Nikolić^{1*},
[0009-0004-3769-6299]

Marina Marjanović²
[0000-0002-9928-6269]

¹The Academy of Applied Technical and
Preschool Studies,
Niš, Serbia

²Singidunum University,
Belgrade, Serbia

Abstract:

This paper explores the emotional connections associated with perfumes by analyzing user reviews and fragrance notes for each product. Using a public dataset sourced from the *Fragnatica* platform, the study applies sentiment analysis techniques to categorize perfumes into six essential emotional groups: Romantic, Energizing, Melancholic, Aggressive, Relaxing, and Neutral. Sentiment analysis models, like VADER, are employed for basic sentiment scoring, while more advanced models including fine-tuned DistilBERT are incorporated to detect nuanced emotions. The emotional tones extracted from user-generated text correlate with consumer ratings and perfume characteristics. The study also investigates the relationship between fragrance notes and user emotions, identifying specific scent profiles that strongly relate to each group. Methodologies applied include sentiment analysis, clustering, and statistical visualization, utilizing a substantial dataset of perfume reviews. These strategies uncover patterns in emotional responses to scent, providing insights into how fragrance compositions influence emotional perceptions. The results bridge the gap between subjective fragrance experiences and objective data analytics, enabling more refined product categorization. Ultimately, this study offers valuable implications for the fragrance industry, helping brands improve product development and marketing strategies by better understanding the emotional resonance, leading to enhanced customer satisfaction and targeted product offerings.

Keywords:

Data Science, Fragrance Notes, Emotion Detection, Sentiment Analysis, Consumer Preferences.

INTRODUCTION

The perfume industry has a significant impact on human emotions, with fragrances playing an important role in shaping moods, memories, and perceptions. Scents are powerful triggers for emotional responses, often evoking feelings of calmness, excitement, nostalgia, or even aggression. This emotional connection between perfumes and individuals is an area of increasing interest, particularly in analyzing how specific fragrances can influence user behavior. Given the complex nature of fragrance experiences, it is crucial to investigate how customers emotionally engage with perfumes and how these emotional associations can be leveraged for better product development and marketing strategies [1].

Correspondence:

Milena Nikolić

e-mail:

milena.nikolic@akademijanis.edu.rs





This paper explores the emotional associations that consumers form with perfumes incorporating popular data science techniques. By analyzing user-generated content from the *Fragnatica* platform, the study applies Natural Language Processing (NLP) models to categorize perfumes into emotional groups such as Romantic, Energizing, Melancholic, Aggressive, Relaxing, and Neutral. Through sentiment analysis, we aim to identify patterns in the emotional responses triggered by different fragrances and how these responses are related to fragrance notes [2]. Moreover, this paper seeks to establish a link between subjective user experiences and data-driven insights, giving actionable recommendations for the fragrance industry.

The following literature review provides a deeper understanding of relevant concepts and methods integrated into this study. The first subsection discusses the role of emotions in fragrance perception and consumer behavior. The second subsection focuses on existing approaches for sentiment analysis, specifically in the context of customer reviews and their applications to fragrance data. Lastly, the third subsection examines the current attempts to determine the relationships between fragrance notes and emotions, exploring how specific scent combinations evoke distinct emotional responses.

1.1. EMOTIONS IN FRAGRANCE PERCEPTION

The study [3] adopts a symbolic perspective to explore consumer demand for "*Scent Library*" perfumes by analyzing online reviews. The research underscores the significance of emotional and social needs in shaping user preferences. Key findings suggest that emotional values such as anticipation, experience, and fun are crucial to customer attraction to perfumes. Additionally, perfumes are a powerful tool for expressing identity and fostering social interactions, supporting brand devotion.

For this research, the authors employed a methodology that collected data from 12,536 valid reviews between January 2023 and January 2024. The data were processed using a Python-based web scraping program, followed by text segmentation and vector mapping for analysis. By exploring reviews from e-commerce platforms, the authors identified 24 evaluation dimensions, which include emotional needs, social identity, and cultural recognition. The findings reveal that emotional needs, particularly anticipation and experience, account for the largest proportion of consumer demand, representing 65% of the reviews.

These results offer perspectives for perfume brands to refine product design and marketing strategies by utilizing emotional and social significance to encourage consumer loyalty. Furthermore, the study highlights the evolving role of perfumes as symbols of culture, identity, and social expression, going beyond their functional purpose.

1.2. SENTIMENT ANALYSIS IN USER REVIEWS

Sentiment analysis plays a significant role in understanding user feedback and making informed decisions in the e-commerce domain. Recent advancements in sentiment analysis models have addressed several challenges, such as capturing the complexity within the context of language and dealing with imbalanced datasets that often result in biased classifications. One such method is the hybrid model proposed in the paper [4], which combines BERT, SMOTE, and VADER. BERT extends the model's ability to understand both local and global dependencies in text, while SMOTE addresses class imbalance by generating synthetic data. VADER further refines the model by providing sentiment labels, achieving an impressive accuracy of 98.1% with real-world data.

Similarly, the use of transformer-based architectures like DistilBERT has shown notable potential in sentiment analysis tasks. According to the study [5], DistilBERT was employed to extract subjective information from customer reviews, categorizing sentiments into different classes, such as positive, neutral, or negative. Through fine-tuning, regularization, and hyperparameter optimization, their approach achieved an accuracy of 86.59% on user reviews, showcasing the strength of transformer-based models for sentiment classification in e-commerce.

By leveraging these models, businesses can gain deeper insights into customer sentiments, helping to elevate product offerings, enhance user experiences, improve marketing strategies, and strengthen brand loyalty.

1.3. COMPUTATIONAL SCENT MODELING

The link between emotions and fragrance perception has been an area of rising research motivation, especially with the integration of artificial intelligence (AI) and natural language processing (NLP) techniques. The emotional impact of fragrances is crucial in consumer preferences, guiding individual choices and marketing strategies.



In the study [6], an NLP-driven framework is introduced for perfume note estimation, using sentence transformer models. The system bonds text descriptions with perfume notes, enhancing recommendation accuracy beyond numerical ratings or basic textual analysis. A main contribution is the *Perfume Notes and Descriptions* dataset, compiled from the *Base Notes* community. The model uses deep learning to predict perfume attributes more accurately. Results show a great improvement in hit rates (37.1%–41.1% to 72.6%–79.0%) and mean reciprocal rank (22.1%–31.9% to 57.3%–63.2%). By fine-tuning transformer models, the study shows the effectiveness of AI in capturing semantic nuances in fragrance descriptions, emphasizing that perception is both chemical and linguistic.

The study [7] also explores the relationship between olfaction, emotions, and esthetics through the development of the *Perfume-Related Olfactory Aesthetic Experience Scale (POLAES)*. This research involved three studies with a total of 677 participants to assess emotions induced by perfumes. Study 1 identified key emotional factors, Study 2 validated them through perfume evaluations, and Study 3 confirmed reliability through test-retest analysis. The final scale consists of 28 items across six factors: Content-Satisfied, Energetic-Romantic, Oblivious of Oneself-Touched, Cold-Aloof, Repulsive-Indifferent, and Desirous-Seduced. Confirmatory factor analysis (CFI = .91) and reliability testing ($r = .83$) demonstrated the scale's validity, reinforcing its usefulness in measuring perfume-related olfactory esthetic experiences.

2. DATA AND METHODOLOGY

In this section, we will provide a detailed overview of the dataset and the methodology used for the analysis. We will describe the data preprocessing steps, which include cleaning the data by handling missing values, removing irrelevant columns, and ensuring the data is ready for further inspection. Following that, we will introduce a BERT-based framework to process perfume notes, descriptions, ratings, and user reviews to determine the emotional category of each perfume. This approach will combine sentiment analysis with a rating-based classification to categorize perfumes into specific emotional groups. By integrating sentiment scoring models like VADER and fine-tuned DistilBERT, we will classify perfumes according to the emotional tones conveyed in reviews and correlate them with their characteristics. This method will provide a better understanding of how fragrance compositions influence emotional perceptions and help us identify scent profiles that strongly align with specific emotional groups.

2.1. DATASET DESCRIPTION

The dataset utilized for this analysis was sourced from *Kaggle*, as a part of the *Fragrantica Data* repository, which consolidates information gathered from the renowned *Fragrantica* platform [8]. This dataset provides a rich set of attributes for a broad range of perfumes, including essential details such as the perfume's name, a description, the designer's name, a list of fragrance notes, a comprehensive collection of reviews, an URL for the perfume's profile, and numerical ratings for certain perfumes.

It is important to highlight that the dataset includes ratings for only a limited subset of perfumes. Specifically, there are 2,474 ratings available from over 84,000 possible values that could have been provided. This disparity in the availability of ratings is a critical consideration, as it means the dataset is not representative of ratings for all perfumes, which could affect the effectiveness of the analysis based on ratings alone. Despite this, the ratings were assigned using an alternative resolution designed to estimate ratings for perfumes where data was lacking. This technique has been elaborated in our previous study [9], where we investigated alternative ways to generate ratings for perfumes based on various features, including fragrance notes and designer popularity.

In addition to addressing the ratings, we performed extensive data cleaning to prepare the dataset for analysis. Specifically, we identified and removed rows where the title or description columns were empty, as these missing values would obstruct any meaningful assessment or machine learning modeling. Furthermore, the URL column, which leads to the perfume's profile on the *Fragrantica* website, was discarded for now. We found that this column was not relevant to the current analysis, as it mainly serves as an external reference in the dataset [10].

The provided Table 1 presents an overview of the key columns in the dataset, along with their descriptions and the number of non-null values available for each column after performing the data preprocessing steps.



Table 1. Overview of the preprocessed dataset

Column name	Data type	Non-null values	Description
title	object	84,136	Name of the perfume.
description	object	84,136	Brief description of the perfume.
designer	object	84,136	Name of the perfume designer.
notes	object	84,136	List of fragrance notes associated with the perfume.
reviews	object	84,136	List of user reviews for the perfume.
rating	float-64	84,136	The average rating of the perfume (1 – 5).

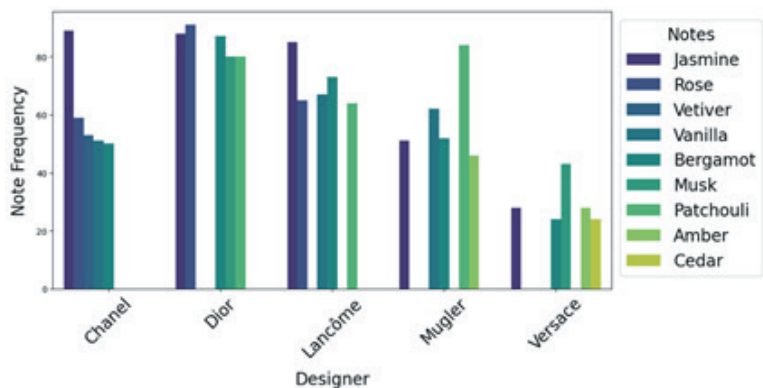


Figure 1. The most common notes by the designer

Through data cleaning steps, we produced a refined dataset that is now well-prepared for deeper processing, including the application of sentiment analysis, emotion-based categorization, and other analytical techniques. With reduced noise and structured information, it enables a more accurate exploration of perfumes and their attributes. Below is an example record from the dataset, showing only one review for better readability.

- **Title:** Classique Wonder Woman Eau Fraiche
- **Description:** Classique Wonder Woman Eau Fraiche by Jean Paul Gaultier is an Amber Floral fragrance for women. Classique Wonder Woman Eau Fraiche was launched in 2017.
- **Designer:** Jean Paul Gaultier
- **Notes:** Sorbet, Ginger, Sugar Cane, Lemon, Orange Blossom, Tiare Flower, Jasmine, Vanilla, Musk, Labdanum
- **Chosen review:** "This must be one of my best blind buys. I bought a 50 ml bottle when I found it in a big sale and thought it would be just a lighter version of JPG Classique that I also own. It does smell very similar to JPG Classique, but it is not quite as soapy, and it is more youthful."
- **Rating:** 4.1

Figure 1 displays the most frequently used fragrance notes across five trending designers (e.g., Chanel, Dior, Versace, etc.), based on their perfume compositions. The data showcases how certain fragrance notes are prevalent in different designer collections, offering details into scent trends and preferences within the industry.

2.2. METHODOLOGY

To effectively categorize perfumes based on their emotional impact, we implemented a comprehensive method that integrates techniques like sentiment analysis, natural language processing (NLP), and statistical correlations between fragrance notes and emotions. By analyzing user-generated reviews, structured numerical ratings, and perfume compositions, we developed a framework that classifies each perfume into one of the six defined emotional categories: Romantic, Energizing, Melancholic, Aggressive, Relaxing, and Neutral.

A key component of this methodology is the analysis of user reviews, which contain rich textual data reflecting individual experiences and emotional responses to different fragrances. To extract meaningful sentiment, we employed VADER (Valence Aware Dictionary and sEntiment Reasoner), a sentiment analysis tool designed



for short and informal text. VADER assigns sentiment scores by analyzing both individual words and contextual modifiers, making it useful for interpreting intricate expressions among perfume reviews. While this strategy offers general sentiment polarity, it does not always capture the deeper emotional layers of the text [11].

Recognizing this limitation, we incorporated a fine-tuned DistilBERT model, a compressed version of BERT, which is a transformer-based deep learning approach that leverages contextual embeddings to identify subtle emotional cues within reviews. DistilBERT, while retaining over 97% of BERT's language comprehension, is more compact and faster, making it particularly well-suited for processing large datasets with improved computational efficiency [12].

In addition to textual sentiment analysis, numerical ratings were introduced to strengthen the classification process. Perfumes with high sentiment scores and strong ratings were predominantly related to positive emotions such as Romantic and Relaxing, whereas those with lower sentiment scores and negative reviews were often linked to emotions such as Aggressive or Melancholic. Neutral sentiment distributions were indicative of either Energizing or Neutral perfumes, depending on the fragrance notes and review content [13].

Beyond sentiment and ratings, we explored the correlation between fragrance notes and emotional perception to establish an empirical relation between scent compositions and emotional responses. By observing the frequency and occurrence of notes across emotional categories, we identified distinct scent-emotion relationships:

- **Romantic** perfumes frequently contain floral (rose, jasmine, tuberose) and fruity (peach, berry, plum) notes, evoking warmth, intimacy, and softness.
- **Energizing** perfumes are predominantly characterized by citrus (lemon, orange, bergamot) and fresh (mint, green tea, aldehydes) accords, commonly associated with vitality and refreshing energy.
- **Melancholic** perfumes feature woody (cedarwood, sandalwood) and smoky (incense, oud) notes, creating a sense of nostalgia, introspection, or depth.
- **Aggressive** perfumes tend to have spicy (black pepper, cinnamon, saffron), leathery, and animalistic (musk, civet, beeswax) compositions, conveying boldness, power, or intensity.
- **Relaxing** perfumes contain aromatic (lavender, chamomile, rosemary), powdery, and vanilla-based accords, promoting a sense of calmness and comfort.
- **Neutral** perfumes do not strongly align with any specific emotional category, but they generally include a balanced blend of multiple accords without a dominant emotional association.

2.3. FINAL CATEGORY EVALUATION

To ensure the credibility of perfume classifications, we applied exploratory data analysis (EDA) and statistical visualization techniques, allowing us to assess the consistency of emotional labels. The final categorization was derived from a composite score value, integrating multiple dimensions of evaluation:

- **Sentiment analysis results**, capturing user-perceived emotional responses in reviews;
- **Numerical ratings**, reinforcing sentiment-based classifications where available; and
- **Note correlations**, identifying recurring patterns between scent compositions and emotional associations.

By combining these elements, we developed a structured framework that balances subjective user experiences with data-driven findings. This approach ensures that each perfume is categorized into one of the six emotional groups in a systematic and reproducible manner, connecting qualitative descriptions and quantitative analysis [14].

Figure 2 illustrates the final categorization model, outlining the complete process for assigning emotional labels to perfumes. It starts with observing sentiment, ratings, and fragrance notes, which are then combined to classify perfumes into their respective emotional categories, ultimately providing the final emotional labels.



3. RESULTS

In this section, we first evaluate the overall distribution of emotional categories within the dataset, which offers a broad understanding of how perfumes are categorized based on emotional tones. This initial step sets the stage for linking individual fragrance notes to specific emotional categories. By observing the prevalence of each emotional category, we gain perspectives of the emotional landscape within the dataset, which provides context for the subsequent analysis of fragrance notes.

Next, we analyze the classification of specific fragrance notes based on their emotional associations. For each note, we examine the perfumes in which it appears and assign it an emotional category based on the dominant category of the perfumes containing that note. This allows us to understand how individual notes contribute to the emotional character of the examined perfumes.

The pie chart in Figure 3 demonstrates the percentage distribution of perfumes across various emotional categories. It reveals that most perfumes in the dataset are melancholic or romantic, while neutral perfumes are the least dominant, as expected. Surprisingly, a significant number of perfumes are classified as relaxing and aggressive.

After categorizing each perfume, as mentioned above, we extended the analysis to individual fragrance notes. For each note, we examined all perfumes in which they appeared and assigned an emotional classification based on the most frequently occurring category among those perfumes.

For instance, vanilla is a note that usually appears in romantic, relaxing, and aggressive perfumes. However, since most perfumes containing vanilla are classified as relaxing, it is mainly associated with calming and soothing emotional qualities. In general, this methodology allows us to understand the emotional tendencies of individual notes based on their widespread occurrence in different perfumes.

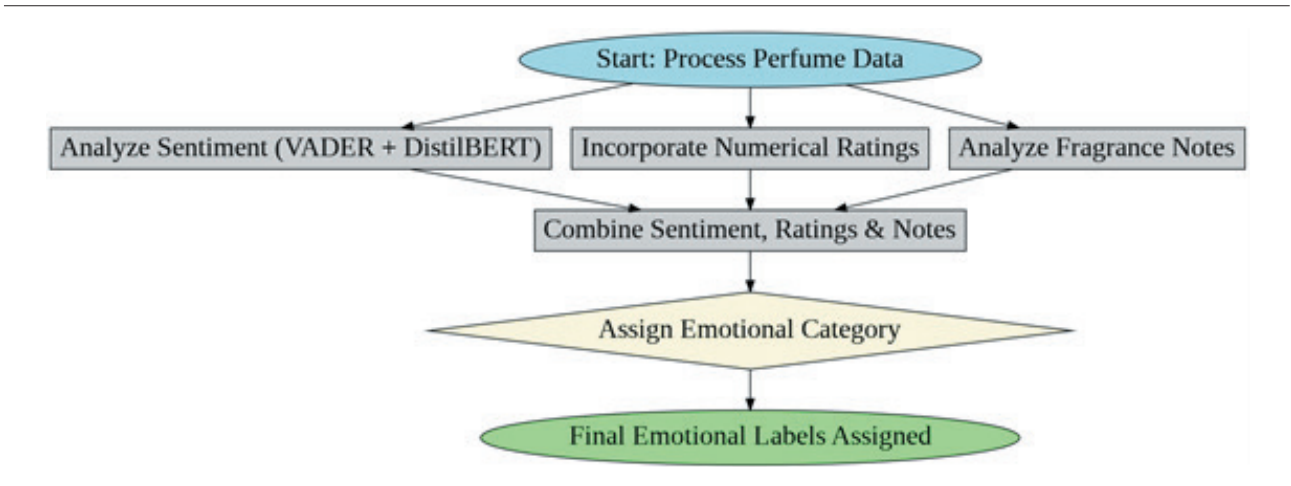


Figure 2. The final categorization model

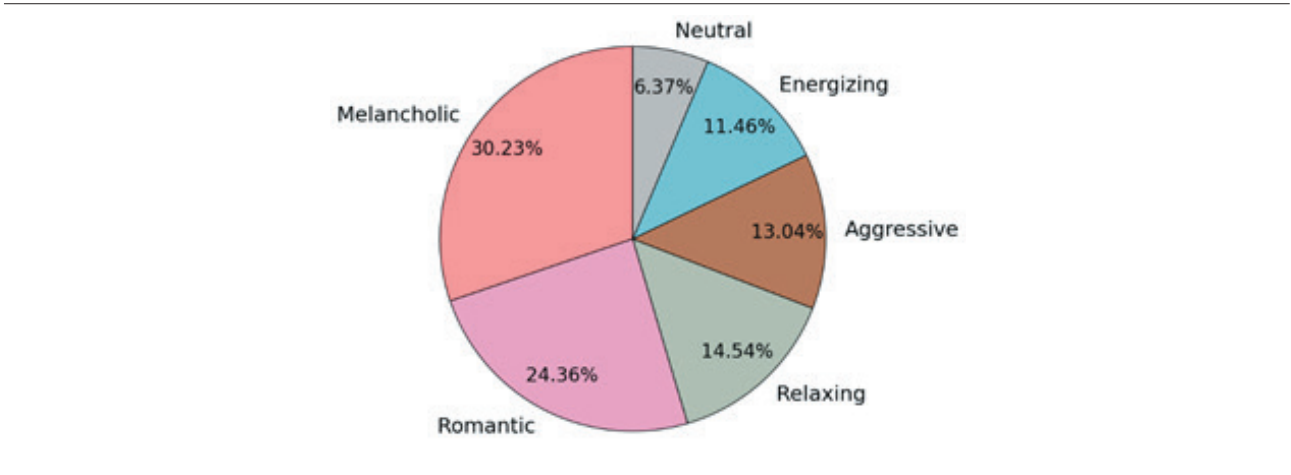


Figure 3. The distribution of emotional categories among perfumes

**Table 2.** Emotional categories, frequency, and brief observations for the top 10 notes

Fragrance notes	Frequency	Emotional category	Observations (This note is present in...)
Musk	31,301	Aggressive	Sensual, bold, long-lasting oriental perfumes.
Bergamot	23,103	Energizing	Fresh, citrusy, uplifting colognes, and summer scents.
Amber	22,693	Romantic	Warm, sweet, resinous oriental, and cozy fragrances.
Sandalwood	22,400	Melancholic	Deep, woody, smooth meditative, and nostalgic scents.
Jasmine	22,153	Romantic	Rich, floral, intoxicating, elegant feminine perfumes.
Patchouli	20,578	Melancholic	Earthy, mysterious, slightly spicy chypre blends.
Vanilla	20,106	Relaxing	Warm, sweet, comforting gourmand, and cozy fragrances.
Rose	18,252	Romantic	Classic, floral, soft and powdery feminine scents.
Cedar	16,130	Melancholic	Woody, dry, masculine, fresh, and clean notes.
Vetiver	11,144	Neutral	Fresh, earthy, grassy, unisex and classic compositions.

To emphasize these findings, Table 2 reveals the top ten most common fragrance notes together with their dominant emotional categories. It also includes key observations derived from the assessment, providing a broader knowledge of the emotional associations linked to each note.

This analysis is crucial for understanding how different fragrance notes resonate emotionally with consumers, providing details into how specific scents evoke feelings. By classifying these notes within emotional categories, we not only gain a more comprehensive awareness of customer preferences but also establish a basis for more accurate and personalized perfume recommendations [15]. These insights are also essential for machine learning models that aim to classify perfumes based on emotional characteristics. By incorporating this data, models can predict scents that match user choices, enhancing product promotion and engagement.

4. CONCLUSION

This research explores the emotional connections consumers make with perfumes, emphasizing how fragrance notes interact with emotional responses. By analyzing user reviews and perfume compositions, we categorized perfumes into distinct emotional groups, like Romantic, Energizing, Melancholic, Aggressive, Relaxing, and Neutral. We then employed sentiment analysis techniques and optimized, pre-trained machine learning models, including a fine-tuned DistilBERT, to extract layered emotional perceptions from the dataset. This combination allowed us to enhance the accuracy of perfume classifications by detecting nuanced emotions within user-generated content.

Through this process, we uncovered strong correlations between fragrance notes and emotional perceptions. For instance, floral and fruity notes were consistently linked to romantic emotions, while citrus and fresh accords evoked energizing feelings. These results offer insights into the fragrance industry, revealing how different scents influence customer emotions and, consequently, purchasing behaviors.

The results of this study have various important implications for the fragrance industry. First, they promote an analytical approach to perfume development, helping beauty brands craft scents that better align with consumer emotional preferences. Second, they provide valuable perspectives for enhancing customer engagement through personalized scent recommendations based on emotional profiles.

Furthermore, the data gathered from reviews and fragrance compositions can be utilized for machine learning models, such as neural networks, to automate the classification process. This would not only boost the accuracy and speed of fragrance categorization but also enable ongoing improvements for the models using real-time data. By incorporating this data into machine learning systems, brands can anticipate trends and create products that better address the emotional needs of users.

In conclusion, the proposed methodology provides a foundation for integrating fragrance experiences with data analytics, opening new doors for targeted product development, personalized consumer engagement, and improved satisfaction in the perfume industry.



5. ACKNOWLEDGMENTS

This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7502, Intelligent Multi-Agent Control and Optimization applied to Green Buildings and Environmental Monitoring Drone Swarms – ECOSwarm.

REFERENCES

- [1] C.-C. Ou and H.-H. Chuang, "Exploring the Factors that Influence Consumers to Purchase Perfume Products," *International Journal of Professional Business Review*, vol. 8, no. 5, 2023.
- [2] J. Mensing, "Psychology of Perfume Choice," in *Beautiful SCENT*, Berlin, Springer, 2023, pp. 53-80.
- [3] W. Li, Y. Jiang and E. Jung, "Exploring Consumer Demand for Perfume Products Based on Symbolic Interaction Theory: A Case Study of "Scent Library" Perfumes," *Psychology Research*, vol. 14, no. 6, pp. 207-213, 2024.
- [4] S. Sharma and N. Desai, "Improvising E-Commerce Sentiment Analysis with Hybrid VADER-BERT Ensemble Model," in *International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, Surabaya, Indonesia, 2024.
- [5] S. Y. Ng, K. M. Lim, C. P. Lee and J. Y. Lim, "Sentiment Analysis using DistilBERT," in *IEEE Conference on Systems, Process & Control (ICSPC)*, Malacca, Malaysia, 2023.
- [6] J. Kim, K. Oh and B.-S. Oh, "An NLP-Based Perfume Note Estimation Based on Descriptive Sentences," *Applied Sciences*, vol. 14, no. 20, p. 9293, 2024.
- [7] Z. Yu, Y. Li and L.-q. Zou, "What perfume makes us feel: A self-report evaluation of perfume-related Olfactory Aesthetic Experience Scale," *Psychology of Aesthetics, Creativity, and the Arts*, 2024.
- [8] J. Hussein, "Kaggle: Fragnatica Data," 2023. [Online]. Available: <https://www.kaggle.com/datasets/joehusseinmama/fragnatica-data>. [Accessed 20 March 2025].
- [9] M. Nikolić, M. Stojanović and M. Marjanović, "Data science techniques for analyzing note trends and brand progression in the perfume industry," in *ISER International Conference*, Prague, Czech Republic, 2024.
- [10] C. Fan, M. Chen, X. Wang, J. Wang and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, 2021.
- [11] S.-H. Park, B.-C. Bae and Y.-G. Cheong, "Emotion Recognition from Text Stories Using an Emotion Embedding Model," in *International Conference on Big Data and Smart Computing (BIGCOMP)*, Kota Kinabalu, Malaysia, 2020.
- [12] H. Wang, X. Kang and F. Ren, "Emotion-Sentence-DistilBERT: A Sentence-BERT-Based Distillation Model for Text Emotion Classification," in *7th International Symposium, ISAIR 2022, Proceedings, Part II*, Shanghai, China, 2022.
- [13] M. Nikolić, M. Marjanović and Ž. Rađenović, "Application of Machine Learning for Perfume Popularity Prediction: Analysis of Fragrance Notes and Consumer Behavior," *Proceedings of the Academy of Applied Technical and Preschool Studies Niš*, pp. 64-67, 2024.
- [14] K. Kalashi, S. Saed and B. Teimourpour, "Sentiment-Driven Community Detection in a Network of Perfume Preferences," *arXiv preprint arXiv:2410.19177*, 2024.
- [15] A. Churchill and J. Behan, "Comparison of methods used to study consumer emotions associated with fragrance," *Food Quality and Preference*, vol. 21, no. 8, pp. 1108-1113, 2010.



APPLICATION OF THE *crAIRsis* AI-BASED FRAMEWORK FOR THE ANALYSIS OF PCB-170 IN HUMAN BREAST MILK

Timea Bezdan¹,
[0000-0001-6938-6974]

Gordana Jovanović²,
[0000-0001-8657-423X]

Andreja Stojić^{1,2},
[0000-0002-5293-9533]

Snežana Herceg Romanić³,
[0000-0003-2382-4734]

Mirjana Perišić^{1,2*}
[0000-0002-8287-4136]

¹Singidunum University,
Belgrade, Serbia

²Institute of Physics Belgrade,
Belgrade, Serbia

³Institute for Medical Research and
Occupational Health,
Zagreb, Croatia

Correspondence:

Mirjana Perišić

e-mail:

mirjana.perisic@ipb.ac.rs

Abstract:

Breast milk is a reliable, non-invasive matrix for monitoring internal exposure to polychlorinated biphenyls (PCBs) and organochlorine pesticides (OCPs), particularly in vulnerable populations such as infants. Ongoing monitoring studies underscore the need for a deeper understanding of the distribution and health impacts of these persistent organic pollutants. Although artificial intelligence (AI) has been widely applied across scientific disciplines, its use in environmental exposure analysis, particularly in biological matrices like breast milk, remains limited. This study investigates the distribution of PCB-170, a highly chlorinated and toxicologically relevant PCB congener, and identifies key predictive factors using an advanced AI-based framework. The analysis was performed using the *crAIRsis* platform, which integrates ensemble machine learning algorithms, metaheuristic optimisation, and explainable AI methods such as Shapley additive explanations (SHAP) and Shapley additive global importance (SAGE). This approach enables the modelling of complex, non-linear relationships between variables. Breast milk samples from 186 mothers in Zadar, Croatia, were analysed for 17 PCB congeners and 7 OCPs. The most influential predictors of PCB-170 levels were PCB-180, PCB-153, and PCB-138, indicating strong co-behaviour and likely shared exposure pathways. These congeners showed relative SHAP impacts ranging from -40% to over 60%. Demographic variables, including maternal age and birth order, had minimal influence, with SHAP impacts below 10%. The study demonstrates the dominant role of higher-chlorinated PCBs in shaping internal burdens and highlights the value of explainable AI in environmental health research. The *crAIRsis* framework offers a robust, transferable methodology for human biomonitoring and evidence-based exposure assessment.

Keywords:

Human Biomonitoring, Polychlorinated Biphenyls, Machine Learning, Metaheuristics, Explainable Artificial Intelligence.

INTRODUCTION

Human milk is a dynamic, bioactive fluid that provides essential nutrients, immune protection, and bioactive compounds crucial for infant development and lifelong health. Its composition varies according to maternal and infant health, diet, and environmental factors, rendering it a form of personalised nutrition. Research into human milk contaminants has expanded since the 1950s, with DDT first detected in 1951 and its metabolites subsequently identified in nearly all tested samples globally. Later surveillance efforts have focused on persistent organic pollutants (POPs), including organochlorine pesticides (OCPs), polychlorinated biphenyls (PCBs), dioxins, organophosphate pesticides,



bisphenols, and polycyclic aromatic hydrocarbons—lipophilic chemicals that accumulate in fatty tissues due to their stability and resistance to degradation. Their presence in breast milk leads to prolonged infant exposure, posing risks such as endocrine disruption, neurodevelopmental effects, and immune dysfunction. As POPs emerged as a global concern, biomonitoring initiatives broadened in scope, revealing temporal trends in legacy pollutants [1] [2] [3]; however, the understanding of pollutant co-occurrence and interrelationships remains limited.

Whilst explainable artificial intelligence (XAI) has been widely applied [4], its integration into environmental research, and its potential to enhance understanding of pollutant dynamics in environmental matrices, remains largely unexplored or frequently misinterpreted. Recently, Huang et al. [5] introduced an innovative approach to assessing chemical exposure risks in breastfeeding infants using an explainable machine learning (ML) model. By integrating ensemble resampling and advanced feature selection techniques, their framework enhances predictive accuracy in identifying high-risk chemicals such as POPs. A key innovation lies in the use of Shapley additive explanations (SHAP), which quantify the contribution of individual features—in this case, chemical properties—to the model's predictions. This approach improves the understanding of molecular factors that influence the transfer of high-risk compounds into human milk and supports more targeted risk mitigation strategies. By identifying molecular fragments linked to high-risk chemicals, the focus shifts from general chemical properties to specific molecular features, thus advancing the field of lactation toxicology. Building on our earlier research [6] [7] [8], which employed ML techniques to investigate dependencies among OCP and PCB congeners in mothers' milk, this study focuses specifically on PCB-170. This highly chlorinated and toxicologically relevant congener has gained attention due to its persistence, bioaccumulative properties, and distinct distribution patterns in environmental and biological matrices.

In this study, we employed a comprehensive AI framework developed within the crAIRsis project [9], which autonomously conducts all stages of the analysis [10] [11]. The objective was to identify the factors influencing PCB-170 distribution in breast milk and evaluate its potential as a predictive marker for broader PCB exposure assessments. The framework integrates seven ensemble regression models selected for their robust predictive power and generalisation performance. In the subsequent phase, each model is systematically evalu-

ated using 25 metaheuristic optimisation algorithms to fine-tune the hyperparameters of the best-performing models and improve predictive accuracy. Model performance is assessed using a suite of evaluation metrics tailored to the specific problem type—classification or regression. Once the optimal model is identified, the framework proceeds to the interpretation phase by incorporating XAI techniques to quantify the contribution of each predictor, both locally and globally, thereby ensuring transparency, interpretability, and actionable insights throughout the analytical workflow. Modelling results are complemented by interactive visualisations that facilitate exploration and interpretation of model behaviour.

Although the primary aim of this paper is to characterise the distribution of PCB-170, the broader ambition of this research is to demonstrate a flexible and robust analytical framework applicable to a wide range of environmental modelling challenges. By combining ML, metaheuristics, and XAI, the proposed approach offers a transferable methodology that supports evidence-based decision-making and strengthens the effectiveness of human biomonitoring strategies.

2. METHODOLOGY

Sample collection and chemical analysis of PCBs and OCPs

Breast milk samples were collected between 2014 and 2019 from 186 healthy mothers (primiparae, secundiparae, and multiparae – third delivery), aged 19 to 41 years, residing in the Zadar region, Croatia. Participants reported no history of accidental or occupational exposure to persistent organic pollutants. Detailed sampling protocols have been described previously [6] [12]. Chemical analysis of PCBs and OCPs followed established procedures outlined in earlier studies [13] [14] [15]. The analysis focused on six indicator PCB congeners (IUPAC numbers: 28, 52, 101, 138, 153, 180), chosen due to their prevalence in technical mixtures, the environment, and biological tissues. Additionally, eleven toxicologically relevant congeners (IUPAC numbers: 60, 74, 105, 114, 118, 123, 156, 157, 167, 189, 170) were included in the analysis.

Data analysis

The data analysis, with PCB-170 as the target variable, was carried out using the crAIRsis framework, a modular and automated AI-based platform that inte-



grates advanced machine learning, metaheuristics, and explainable artificial intelligence techniques. At its core, the framework incorporates seven ensemble regression algorithms: AdaBoost, CatBoost, ExtraTrees, Gradient Boosting, Histogram Gradient Boosting, LightGBM, and XGBoost [16] [17] [18] [19]. These algorithms were selected for their proven ability to capture complex, non-linear patterns and their robustness against overfitting. Each model was evaluated using five-fold cross-validation, ensuring generalisability and minimising bias. Based on the evaluation metrics—specifically R-squared, mean absolute error, mean squared error, and others—the three best-performing models were selected for further optimisation. To enhance predictive accuracy, the hyperparameters of these top models were fine-tuned using the Sine Cosine Algorithm and Harris Hawks Optimisation metaheuristic methods [20] [21]. These approaches efficiently explore the hyperparameter search space and have demonstrated strong performance in solving complex, non-convex optimisation problems in machine learning contexts. Once the final model was selected, explainability and interpretation were undertaken using SHAP and SAGE. SHAP quantifies the influence of each input feature on individual predictions, providing a detailed understanding of model behaviour at the instance level [22], whereas SAGE evaluates feature importance by aggregating their contributions across the entire dataset, thereby identifying variables with the strongest overall impact on model performance [23]. To support interpretation, we further derived relative and normalised SHAP values and introduced a categorical framework referred to as inherent SHAP values. Relative SHAP values express the proportion of a feature's absolute SHAP value in relation to the total attribution for a given prediction, offering insight into the feature's contextual importance. Normalised SHAP values, scaled to the expected model output, simplify the interpretation of impact magnitudes. Inherent SHAP values were used to group SHAP effects into interpretable categories, whereby high negative impacts were defined as those falling below the mean of all negative SHAP values. All these methods contribute to enhanced model transparency and facilitate the interpretability of complex relationships between input variables and the target outcome. To gain a deeper understanding of feature interactions and group-level behaviour, cluster analysis was applied to the SHAP values. Dimensionality reduction was performed using Pairwise Controlled Manifold Approximation [24], followed by clustering using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [25] [26]. This

stage enabled the identification of patterns, subgroups, and outliers within the dataset based on model behaviour.

The crAIRsis framework automates all key stages, including model training, evaluation, optimisation, explanation, and visualisation of both raw data and results, offering a robust and transferable methodology for environmental data modelling. All outputs, including evaluation metrics, feature contributions, and visualisations, are automatically generated and structured for subsequent interpretation and reporting.

3. RESULTS AND DISCUSSION

Monitoring of POPs in human milk in Croatia dates back to the 1970s, primarily focusing on the assessment of PCB and OCP levels and their temporal trends. Our previous studies and review [2] [12], centred mainly on monitoring efforts, revealed a consistent decline in p,p'-DDE, HCB, β -HCH, and PCBs from 1976 to 2019, largely attributed to regulatory restrictions introduced under the Stockholm Convention since the 1990s. As previously reported [2] [12], within the dataset used for this study, compounds from the DDT group—particularly p,p'-DDE—were the most prevalent, followed by PCB-153, PCB-138, PCB-180, β -HCH, PCB-118, γ -HCH, HCB, PCB-156, and PCB-170, in decreasing order. The least abundant compounds included PCB-28, PCB-105, and PCB-60. Detailed descriptive statistics and comparisons with other studies have also been documented previously.

Our pioneering research, aimed at gaining a more precise understanding of the levels, interrelations, and associations of PCBs and OCPs in human milk with maternal factors such as age and parity, employed machine learning techniques. In Jovanović et al. [6], the Guided Regularised Random Forest (GRRF) algorithm identified key factors influencing POP levels, revealing strong non-linear relationships among pollutants and the complexity of their pathways in breast milk. The model achieved prediction errors below 30% and correlation coefficients exceeding 0.90 between predicted and observed values.

Building on this, Jovanović et al. [8] applied advanced machine learning methods—XGBoost and SHAP—to investigate PCB-138 interactions with other non-dioxin-like congeners, maternal age, and parity, identifying PCB-170 and PCB-153 as key drivers of PCB-138 behaviour in milk samples and promising targets for further research.



The ExtraTrees model, optimised using the Sine Cosine Algorithm (SCA), demonstrated strong predictive performance for PCB-170, achieving an R^2 of 0.9125, RMSE of 0.7354, and MAPE of 0.3234, indicating high accuracy and a substantial proportion of explained variance in predicted concentrations. The most influential variables in predicting PCB-170 concentrations were PCB-180, PCB-153, and PCB-138, all of which exhibited significantly higher importance scores than other predictors (Figure 1). These three congeners dominated the model's output, indicating strong associations and potential co-behaviour patterns with PCB-170. While this may reflect similarities in chemical structure and possible overlap in exposure sources or metabolic pathways, such patterns do not necessarily imply a shared origin or an increased tendency for bioaccumulation. Rather, the observed co-behaviour suggests that these congeners may be influenced by related biological or environmental factors that shape their distribution in human milk [27][28]. Among the other variables, PCB-156, p,p'-DDE, and PCB-118 also contributed to the model, though with notably lower impact values. In contrast, factors such as birth order, maternal age, and certain OCPs (e.g. β -HCH, p,p'-DDT) showed minimal influence, suggesting a limited role in explaining PCB-170 variability within the analysed dataset.

Figure 2 presents SHAP dependence plots illustrating the relative impact of PCB-180, PCB-153, and PCB-138 on the prediction of PCB-170 concentrations. By applying a clustering approach to the SHAP values, we identified distinct groups of variable impacts, each representing a specific cluster that highlights how key factors influence the biological and exposure-related

context affecting PCB-170 levels. These clusters are shown in different colors on the plot, and individual samples (represented as dots) are colored according to the cluster to which they belong. A more in-depth exploration of these patterns lies beyond the scope of this study. Across all plots, a clear monotonic relationship is observed between each specific congener and PCB-170, with higher concentrations of each predictor corresponding to an increased contribution to the predicted PCB-170 level. The strongest relative impacts, up to 60% for PCB-138 and PCB-153, and up to 40% for PCB-180, are associated with mid-to-high concentration ranges of PCB-153 and PCB-180, suggesting that these congeners play a dominant role in shaping the modelled PCB-170 dynamics. Marginal effects remain stable within low-concentration ranges but exhibit greater variance and influence as values increase, particularly for PCB-153. This indicates that interactions with other features are not fully disentangled, which could potentially be addressed by expanding the dataset, both in terms of sample size and inclusion of additional relevant variables. These patterns further support the notion of strong co-behaviour and likely shared exposure pathways among higher-chlorinated PCBs in the analysed population. Moreover, in the lower concentration range of the dominant predictors, negative contributions to PCB-170 prediction are also observed (with relative impacts of up to -40%), indicating that, in certain breast milk samples, the co-occurrence of these compounds at observed concentration levels is associated with reduced PCB-170 levels. This may reflect competitive metabolic pathways, differential accumulation patterns, or individual differences in exposure and elimination dynamics.

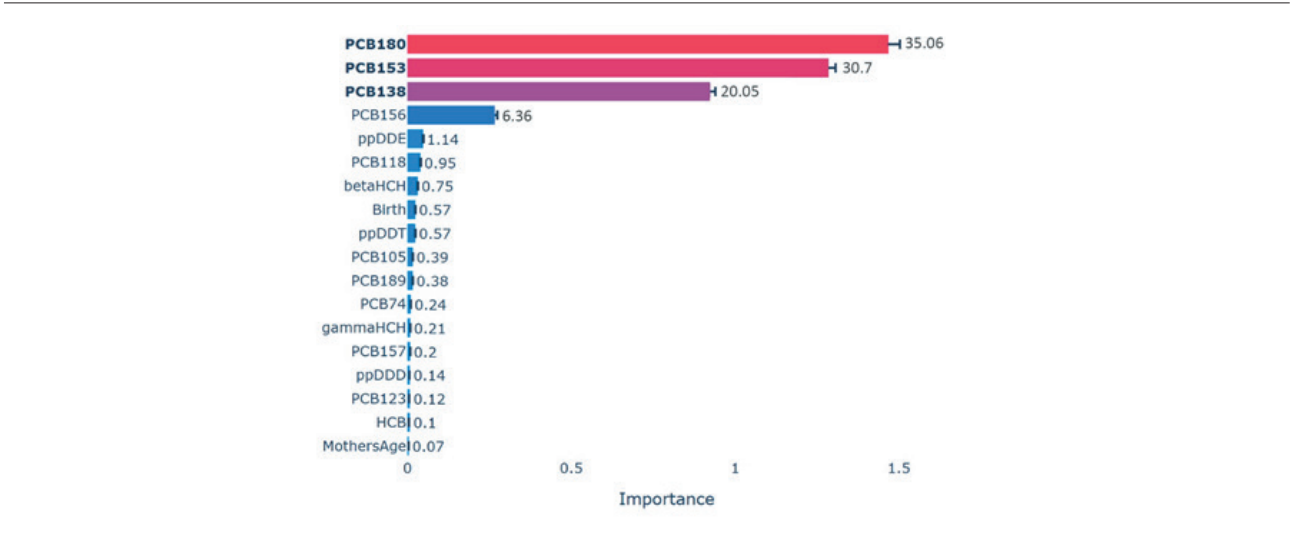


Figure 1. Global feature importance based on SAGE values. Normalized SAGE values [%] were provided as numbers within or next to the bars



The analysis indicates that maternal age has minimal influence, with SHAP values distributed closely around zero across the entire age range (Figure 3, upper panel), suggesting that, within the present dataset, age does not significantly contribute to the prediction of PCB-170 levels. In contrast, birth order exhibits slightly greater variation in its impact, particularly among mothers with three children (Figure 3, lower panel). Although the overall relative influence remains modest, there is a slight tendency towards higher predicted PCB-170 concentrations in multiparous women. Older mothers are generally subject to greater cumulative exposure to POPs, whereas parity is often associated with lower POP levels, as breastfeeding facilitates the reduction of maternal body burdens. This was demonstrated in a

Norwegian study, where age was positively associated with POP concentrations, while increased parity correlated with lower levels [29]. More recent research, including a study involving over one thousand primiparous women in the Czech Republic, found that maternal diet before and during pregnancy influenced PCB levels in breast milk, whereas body weight and age at delivery had no significant effect [27]. Nevertheless, the effect size of maternal age and parity is substantially smaller than that of chemical predictors related to concentration levels, suggesting their limited apparent role in shaping internal PCB-170 burden. This likely reflects insufficient characterisation of maternal status, highlighting the need for additional variables that capture physiological and exposure-related factors more comprehensively.

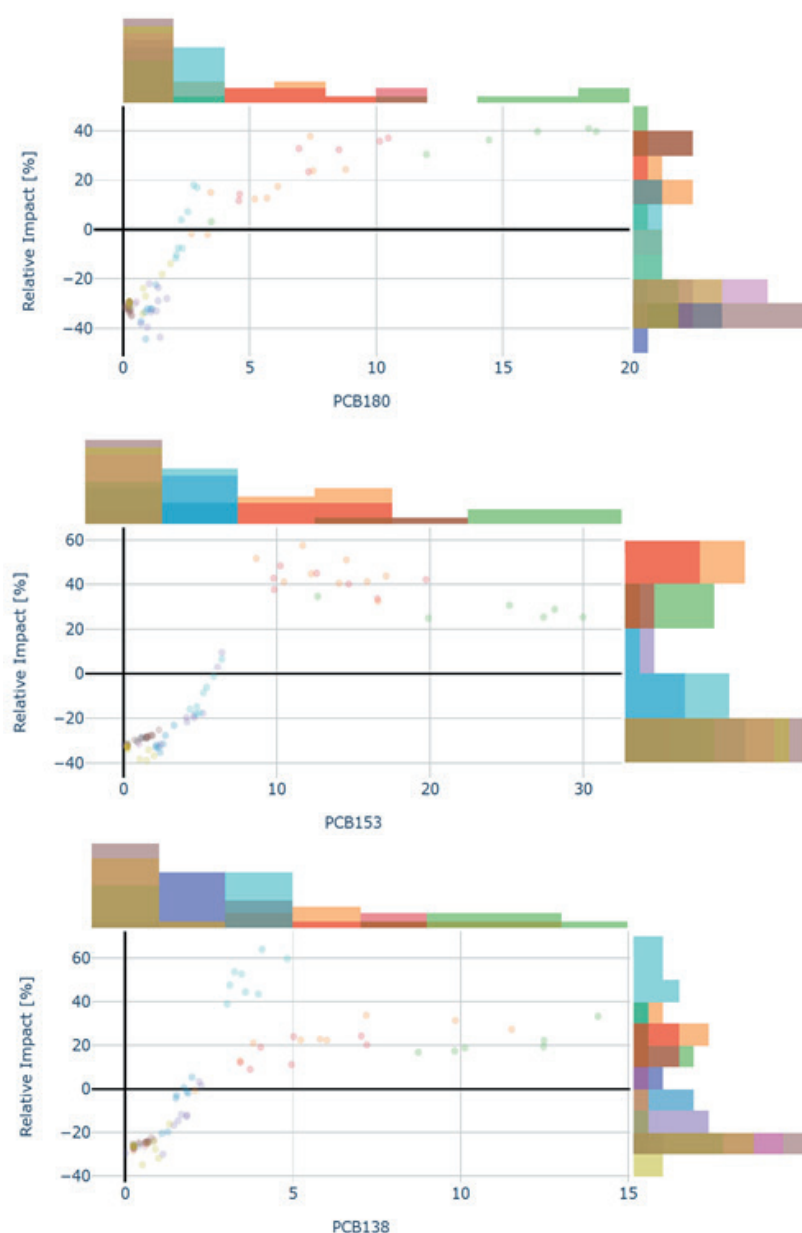


Figure 2. Relative SHAP impacts of the most important variables PCB-180, PCB-153 and PCB-138 on PCB-170 dynamic

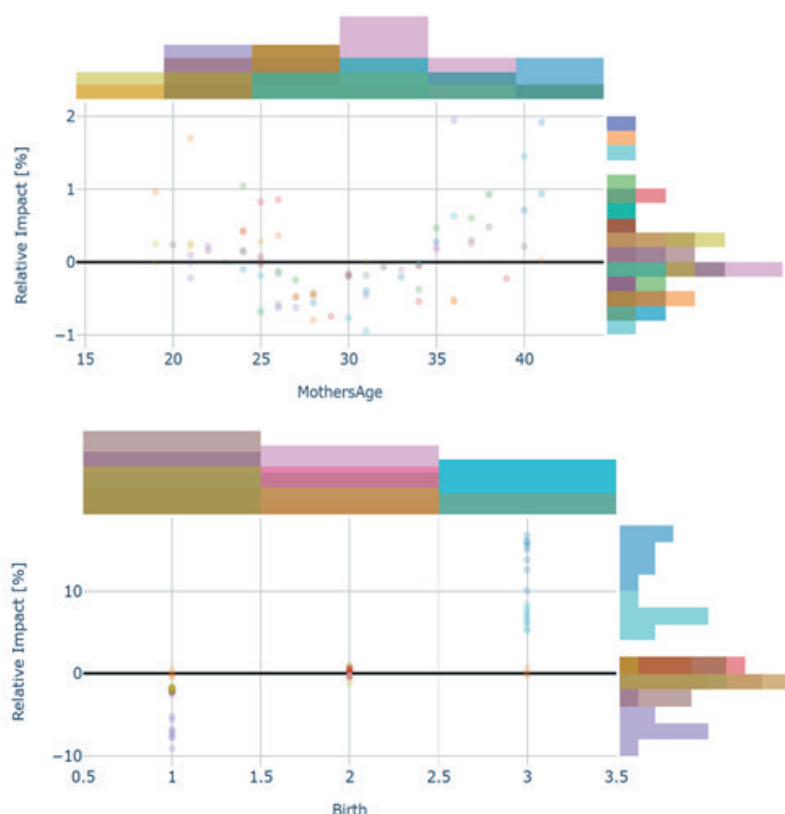


Figure 3. Relative SHAP impacts of the mothers age and birth order on PCB-170 dynamic

4. CONCLUSION

This study demonstrated the effectiveness of the crAIRsis AI-based framework in modelling and enhancing the understanding of PCB-170 distribution in human breast milk. By combining ensemble machine learning algorithms, metaheuristic optimisation, and explainable AI, the framework provided more transparent, data-driven insights into the complex, non-linear relationships between PCB-170 and its predictors. The results identified PCB-180, PCB-153, and PCB-138 as the most influential variables, indicating strong co-behaviour and likely shared exposure pathways with PCB-170. Moderate contributions were observed from other chemical compounds such as PCB-156 and p,p'-DDE, while demographic factors, maternal age and birth order, exerted minimal influence, suggesting that direct chemical exposures play a dominant role in shaping PCB-170 concentrations. Beyond supporting the role of higher-chlorinated congeners, this work illustrates the potential of an automated and interpretable framework for environmental data analysis and human biomonitoring. Future research will build on these findings by conducting a more detailed evaluation of all included variables, exploring potential interactions and non-linear effects. In addition, clustering of SHAP values will be

systematically integrated into the interpretation phase to identify subpopulations with distinct exposure profiles. Finally, the methodology offers a transferable tool for evaluating exposure to persistent organic pollutants and supporting evidence-based health risk assessments.

5. ACKNOWLEDGEMENTS

The authors acknowledge funding provided by the Institute of Physics Belgrade through a grant from the Ministry of Education, Science and Technological Development of the Republic of Serbia, as well as by the Science Fund of the Republic of Serbia (Grant No. 7373, Characterizing crises-caused air pollution alterations using an artificial intelligence-based framework – crAIRsis). The analysis of PCBs and OCPs was carried out using facilities and equipment funded by the European Regional Development Fund, under the project KK.01.1.1.02.0007 "Research and Education Centre of Environmental Health and Radiation Protection – Reconstruction and Expansion of the Institute for Medical Research and Occupational Health", and co-funded by the European Union – Next Generation EU (project EnvironPollutHealth, Program Contract dated 8 December 2023, Class: 643-02/23-01/00016, Reg. No. 533-03-23-0006).



REFERENCES

- [1] J. Fång, "Spatial and temporal trends of the Stockholm Convention POPs in mothers' milk — a global review," *Environ. Sci. Pollut. Res.*, p. 8989–9041, 2015.
- [2] S. H. Romanić, "Persistent organic pollutants in Croatian breast milk: An overview of pollutant levels and infant health risk assessment from 1976 to the present", *Food. Chem. Toxicol.*, p. 113990, 2023.
- [3] K. R. Nermo, "Trend analyses of persistent organic pollutants in human milk from first-time mothers in Norway between 2002 and 2021," *Int. J. Hyg. Environ.*, p. 114458, 2025.
- [4] D. Gunning, "XAI—Explainable artificial intelligence," *Sci. Robot.*, p. 7120, 2019.
- [5] X. Huang, "Assessing chemical exposure risk in breastfeeding infants: An explainable machine learning model for human milk transfer prediction," *Ecotoxicol. Environ. Saf.*, p. 117707, 2025.
- [6] G. Jovanović, "Introducing of modeling techniques in the research of POPs in breast milk – A pilot study," *Ecotoxicol. Environ. Saf.*, pp. 341–347, 2019.
- [7] A. Stojić, "Shapley additive explanations of indicator pcb-138 distribution in breast milk," in *international scientific conference on Information technology and data related research - Sinteza, Singidunum University*, Belgrade, Serbia, Belgrade, 2020.
- [8] G. Jovanović, "Patterns of PCB-138 Occurrence in the Breast Milk of Primiparae and Multiparae Using SHapley Additive exPlanations Analysis," in *Artificial Intelligence: Theory and Applications*, Springer, Cham, 2021, pp. 191–206.
- [9] "Project crAIRsis – Characterizing Crises-Caused Air Pollution Alterations Using an Artificial Intelligence-Based Framework," cience Fund of the Republic of Serbia (Grant No. 7373), PRISMA program, 2024–2027.
- [10] G. Jovanović, "The PM 2.5-Bound Polycyclic Aromatic Hydrocarbon Behavior in Indoor and Outdoor Environments, Part III: Role of Environmental Settings in Elevating Indoor Concentrations of Benzo (a) pyrene," *Atmosphere*, 2024.
- [11] A. Stojić, "Artificial Intelligence-Based Framework for Analyzing Crises-Caused Air Pollution," in *Sinteza 2024-International Scientific Conference on Information Technology, Computer Science, and Data Science. Singidunum University*, Belgrade, 2024.
- [12] G. Mendaš, "Presence of polycyclic aromatic hydrocarbons and persistent organochlorine pollutants in human Milk: Evaluating their levels, association with Total antioxidant capacity, and risk assessment," *Sci. Total Environ.*, p. 172911, 2024.
- [13] D. Klinčić, "Polychlorinated biphenyls and organochlorine pesticides in human milk samples from two regions in Croatia," *Environ. Toxicol. Pharmacol.*, pp. 543–552, 2014.
- [14] D. Klinčić, "Organochlorine pesticides and PCBs (including dl-PCBs) in human milk samples collected from multiparae from Croatia and comparison with primiparae," *Environ. Toxicol. Pharmacol.*, pp. 74–79, 2016.
- [15] I. Šimić, "Optimization of gas chromatography-electron ionization-tandem mass spectrometry for determining toxic non-ortho polychlorinated biphenyls in breast milk," *Biomed. Environ. Sci.*, pp. 58–61, 2020.
- [16] Y. Freund, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, p. 119–139, 1997.
- [17] L. Prokhorenkova, "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
- [18] G. Ke, "Lightgbm: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.*, p. 1189–1232, 2001.
- [20] X.-S. Yang, "Firefly Algorithms for Multimodal Optimization," in *Stochastic Algorithms: Foundations and Applications*, O. Watanabe and T. Zeugmann, Eds., Springer Berlin Heidelberg, 2009, pp. 169–178.
- [21] A. A. Heidari, "Harris hawks optimization: Algorithm and applications," *Future Gener. Comput. Syst.*, pp. 849–872, 2019.
- [22] S. M. Lundberg, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [23] I. Covert, "Understanding global feature contributions with additive importance measures," *Adv. Neural Inf. Process. Syst.*, pp. 17212–17223, 2020.
- [24] Y. Wang, "Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization," *Journal of Machine Learning Research*, pp. 1–73, 2021.
- [25] L. McInnes, "Umap: Uniform manifold approximation and projection for dimension reduction," *ArXiv Prepr.*, p. 180203426, 2018.
- [26] L. McInnes, "HDBSCAN: Hierarchical density based clustering," *J Open Source Softw*, p. 205, 2017.
- [27] R. Aerts, "Determinants of persistent organic pollutant (POP) concentrations in human breast milk of a cross-sectional sample of primiparous mothers in Belgium," *Environ. Int.*, p. 104979, 2019.
- [28] J. Komprda, "Dynamics of PCB exposure in the past 50 years and recent high concentrations in human breast milk: analysis of influencing factors using a physiologically based pharmacokinetic model," *Sci. Total Environ*, pp. 388–399, 2019.
- [29] A. Polder, "Levels of chlorinated pesticides and polychlorinated biphenyls in Norwegian breast milk (2002–2006), and factors that may predict the level of contamination," *Sci. Total Environ.*, pp. 4584–4590, 2009.



THE MANN-KENDALL TEST FOR WEATHER FORECASTING IN SMART CITIES

Miloš Todorov^{1*},
[0000-0001-6938-6974]

Ninoslava Tihi²,
[0000-0001-8657-423X]

Marko Pavlović³,
[0000-0002-5293-9533]

Suad Suljević³
[0000-0003-2382-4734]

¹The Faculty of Mathematics and
Computer Science, Alfa BK University,
Belgrade, Serbia

²The Higher Education Technical School
of Professional Studies,
University of Novi Sad,
Novi Sad, Serbia

³MB University,
Belgrade, Serbia

Correspondence:

Miloš Todorov

e-mail:

milos.todorov@alfa.edu.rs

Abstract:

In smart cities, predicting and responding to the coming of severe weather is one of the most important advantages that can be gained from precise weather forecasting. As a direct consequence of climate change, the planet is experiencing an increased frequency of heatwaves, storms, and floods more often than in the past. To be able to respond promptly, local government agencies must have access to precise weather predictions, which would help them in decision making and effectively delivering emergency aid, and, if necessary, evacuating areas that are at a larger risk of danger. In addition to this, people can use mobile apps to get weather forecasts in a short amount of time. This article employs the Mann-Kendall test to analyse monthly time series data, specifically focusing on weather parameters like temperature, soil moisture, and precipitation. This test will certainly provide useful insights to choose the best appropriate model for predicting drought-related parameters, which will be later visualized and presented to decision makers.

Keywords:

Statistical Analysis, The Mann-Kendall Test, Weather Forecasting, Smart Cities.

INTRODUCTION

Time series forecasting is a statistical technique used to make predictions about future values of a time-dependent data set based on past observations [1]. It is commonly used in various fields such as economics, finance, and meteorology to forecast trends and make informed decisions. To integrate the prediction accuracies of machine learning methodologies, the framework for forecasting precipitation in smart cities utilises fuzzy logic. This procedure, which is referred to as fusion, was proposed by Atta-ur Rahman [2]. In Europe, the effects of soil moisture on seasonal temperature and precipitation prediction scores have been shown by Van Den Hurk [3].

One popular method for time series analysis is the Mann-Kendall test, which is a non-parametric test used to detect trends in time series data [4], [5]. This test analyses the data for monotonic trends, which are trends that consistently increase or decrease over time. The non-parametric Mann-Kendall test is suggested due to its robustness and lack of assumptions on data distribution.



It is used to ascertain the direction of the trend in a time series, whether it is upward or downward. It is a process that uses ranks instead of raw data values. It is suitable for data that does not have a normal distribution, has outliers or exhibits non-linear trends [4],[5]. This methodology is often used to estimate trends in a time series of mean, high and low flows, capturing their temporal variations and facilitating the prediction of future trends.

Another important technique in time series analysis is the Theil-Sen trend method, which is a robust method for estimating trends in time series data. This method is resistant to outliers and does not assume a specific distribution of the data, making it a reliable option for analysing time series data with irregular patterns or anomalies. The Theil-Sen trend method calculates the median of all possible slopes between data points, providing a robust estimate of the trend without being influenced by extreme values. Overall, time series trend analysis techniques, such as the Mann-Kendall test and Theil-Sen trend method, play a crucial role in analysing and predicting future trends in time-dependent data sets, helping researchers and decision-makers make informed decisions based on historical patterns and trends [6].

In addition, the Theil-Sen estimator is often used as a method for calculating trend slope often in combination with other trend analysis techniques, particularly the Mann-Kendall test [4], [5], [6], [7]. The statistically significant trends are determined using the Mann-Kendall test at a certain level of significance. The trend quantification and extension are thereafter carried out using the Theil-Sen estimator. This estimator is regarded as a reliable mathematical tool for estimating the extent of a trend. It does this by calculating the median value over all possible combinations of pairings within the given time series [7].

The Mann-Kendall test is particularly useful for identifying trends in time series data that may not be linear or follow a specific pattern. It is widely used in environmental studies to detect changes in variables such as precipitation patterns, temperature fluctuations and air pollution levels. The test calculates the Kendall rank correlation coefficient, which measures the strength and direction of any trend present in the data. A positive coefficient indicates an increasing trend, while a negative coefficient indicates a decreasing trend [4], [5].

The Theil-Sen estimator is a commonly used methodology for determining the slope of a trend, generally employed in combination with other methods of trend research, particularly the Mann-Kendall test[6]. The Mann-Kendall test and the Theil-Sen estimator are often used in many applications to estimate gradual trends, which are then utilized in the forecast of different weather parameters [8], [9], [10], [11].

In this paper, the application of the Mann Kendall test on soil moisture, precipitation and temperature will be discussed.

2. THE MANN- KENDALL TEST AND THEIL-SEN TREND ESTIMATOR

In this study, a two-stage technique is used to estimate trends. The first phase involves applying the non-parametric Mann-Kendall test to find trends. The second step involves modelling the trends using the Theil-Sen trend estimator. Figure illustrates this process.

Therefore, a two-step procedure for trend estimation is applied at the beginning of this research:

- (Step 1) Identifying trends with the the Mann-Kendall test;
- (Step 2) Modelling trends using the Theil-Sen trend estimator.

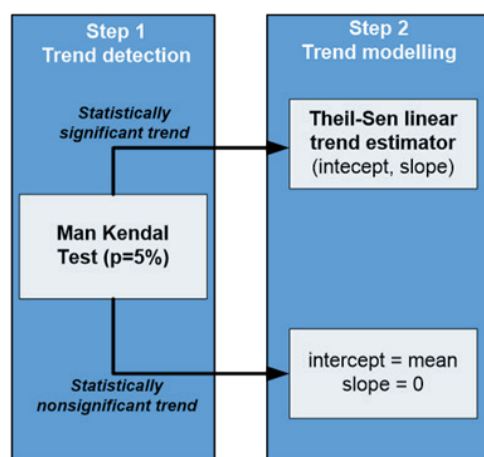


Figure 1. A chart diagram for the trend detection in modelling the temperature time series



Step 1 – The Mann Kendall test

The Mann-Kendall trend test utilizes Kendall's statistic (S) to score each data point throughout the whole time series [4],[5], [12]:

$$S = \sum_{j=1}^{n-1} \sum_{i=j+1}^n \text{sign}(Q_i - Q_j) \quad (1)$$

where n represents the total number of time series elements and Q denotes the mean monthly or annual flow at time increments i and j , respectively. Furthermore, the value of $\text{sign}()$ is determined by whether the time series member at time step i is greater than the corresponding member at time step j . If so, $\text{sign}()$ returns +1; otherwise, it returns -1. If the time series member at time step i is equal to the corresponding member at time step j , $\text{sign}()$ returns 0.

Variance of Kendall's statistics $\text{Var}(S)$ is calculated as [4], [5], [12]:

$$\text{Var}(S) = \begin{cases} \frac{1}{18} [n \cdot (n-1) \cdot (2n+5)], & \text{if } n = l \\ \frac{1}{18} [n \cdot (n-1) \cdot (2n+5) - \sum_{q=1}^l t_q \cdot (t_q-1) \cdot (2t_q+5)], & \text{else} \end{cases} \quad (2)$$

where n is a total number of time series members, l is number of groups which represent the set of sample data with the same values and t_q is number of observations (frequencies) in the q^{th} group.

Once Kendall statistic is determined, the Mann-Kendall test statistic (Z) for the time series is calculated as follows [4],[5], [12]:

$$Z = \begin{cases} \frac{S-1}{\sqrt{\text{Var}(S)}}, & \text{if } S > 0 \\ 0, & \text{if } S = 0 \\ \frac{S+1}{\sqrt{\text{Var}(S)}}, & \text{if } S < 0 \end{cases} \quad (3)$$

Since the Mann-Kendall statistic (Z) conforms to a normal distribution, the probability density function for a normal distribution $N(0,1)$ is defined as:

$$f(Z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{Z^2}{2}} \quad (4)$$

The null hypothesis is accepted when the probability value p is greater than 0.05 and there is no discernible trend in the time series. Conversely, at a significance level of $p \leq 0.05$ the alternative hypothesis is accepted indicating the presence of a trend in the time series which may be negative or positive.

Step 2 – The Theil-Sen trend estimator

If the Mann-Kendall test statistic (Step 1) suggests that a substantial trend continues to exist in the time series, the Theil-Sen trend estimator is utilized to calculate the linear trend parameters including the intercept and slope. Consequently, the extent of the linear trend in time series data is determined by averaging the pairwise slopes [6], [12].

$$f(Z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{Z^2}{2}} \quad (5)$$

However, when there are statistically non-significant trends, there is no need to use the Theil-Sen trend estimator since the linear trend slope is 0 (Figure).

```
#Importing libraries
import numpy as np
import pandas as pd
import pymannkendall as mk

#Reading datasets
padavine=pd.read_excel('padavine.xlsx')
padavine.head()
temperatura=pd.read_excel('temperatura.xlsx')
temperatura.head()
vlaznost_zem=pd.read_excel('vlaznost_zem.xlsx')
vlaznost_zem.head()

#Applying Mann-Kendall test
padavineTest = mk.original_test(padavine['Vrednost'])
print(padavineTest)
temperaturaTest = mk.original_test(temperatura['Vrednost'])
print(temperaturaTest)
vlaznost_zemTest = mk.original_test(vlaznost_zem['Vrednost'])
print(vlaznost_zemTest)
```

Listing 1. A Python program for applying the Mann-Kendall test



3. DATASET AND FRAMEWORK FLOW

Data were gathered from a weather station located near Novi Sad from January 2014 to December 2020. The data collected by this network station comes from three sensors that measure the temperature of the air, the amount of precipitation, and the moisture content of the soil. Every hour of the day, measurements are carried out.

Following the retrieval of the data in Excel format from the web portal, the data are transformed and then stored in a tabular data structure called data frame for further manipulation. The preparation process continues with filtering and aggregating using the "mean" and "sum" functions to acquire the monthly average values and finally sorting the obtained data by months and years and exporting them in an Excel file.

3.1. CODE ALGORITHM IN PYTHON

Several Python programming modules that are suited for this research were used to construct specialised scripts that were produced expressly for the aim of this study. It was through the use of the Python program script that integrates a few Python packages that the outcomes were obtained [12].

Listing 1 shows an example of a small program written in the Python programming language. It contains importing libraries, reading dataset and applying the Mann-Kendall test over loaded data.

4. RESULTS AND ANALYSIS

The following table presents the outcomes derived from the Python script.

The Mann-Kendall test provide a combination that includes the following components:

- **Trend:** indicates the trend (whether it is rising, decreasing, or there is no trend at all);
- **p-value:** p-value of the significance test;
- **Slope:** Theil-Sen estimator/slope;

- **Intercept:** the intercept of the Kendall-Theil Robust Line, for the seasonal test, with the whole period cycle regarded as a unit time step.

As shown in Table 1, trend analysis provides results for three meteorological variables: temperature, precipitation and soil moisture. No significant trend is present in the temperature time series according to the obtained p value from the Mann-Kendal test ($p=0.6569$), meaning there's no clear increasing or decreasing pattern. No significant trend has also been found in the precipitation time series ($p=0.1529$) and therefore it can be concluded that this time series is relatively stable over time. However, in the soil moisture time series the Mann-Kendall test has detected a decreasing trend with a highly significant p value ($p=3.28 \times 10^{-5}$), indicating a real downward trend in the soil moisture levels. The slope represents the rate of change of each variable over time. The temperature and precipitation time series have a slight increase but not very significant slope, but the soil moisture has a significant decrease meaning the soil moisture is consistently declining over time. The intercept represents the expected value of the variable at the start of the dataset. Temperature started around 12.37 °C, precipitation started around 39.05 mm while soil moisture started around 25.54% but it is decreasing significantly.

5. CONCLUSION

Smart cities benefit most from accurate weather forecasting because they can prepare for severe weather. Heatwaves, storms, and floods have risen globally owing to climate change. Local government agencies can plan, give emergency help, and evacuate high-risk regions with accurate weather forecasts. Mobile applications can quickly provide weather predictions, allowing homeowners to change their plans and take safety precautions.

Accurate weather forecasting in intelligent cities goes beyond crisis management. Transportation, tourism, and agriculture depend on accurate weather predictions. Agricultural professionals can predict crop yields, control irrigation, and safeguard crops from weather.

Table 1. Output results of applying the Mann-Kendall test in Python code

Time Series	Trend	p-value	Slope	Intercept
Temperature	No trend	0.6568582113023143	0.0138589659774	12.36958549592
Precipitation	No trend	0.1528558365907872	0.1939764492753	39.05002264492
Soil moisture	Decreasing	3.284048567153697e-05	0.0627024599945	25.53967055026



By optimising routes and schedules, transport firms can reduce delays and improve customer satisfaction. Being aware helps people plan for outside events and activities, reducing weather-related difficulties.

The Mann-Kendall test is an extremely useful instrument that can be used to recognise monotonic patterns in time series data. This makes it very pertinent for the field of environmental research. When it comes to estimating trends in time series data, the Theil-Sen trend technique offers a straightforward and trustworthy methodology. After conducting the tests, the lack of a trend in the temperature and precipitation time series suggests that other external factors (e.g., local climate variability) may be influencing them. The decrease in soil moisture could indicate increasing drought conditions, reduced groundwater recharge, or changes in land use affecting water retention.

REFERENCES

- [1] K. C. Tan, 'Trends of rainfall regime in Peninsular Malaysia during northeast and southwest monsoons', *J. Phys. Conf. Ser.*, vol. 995, p. 012122, Apr. 2018, doi: 10.1088/1742-6596/995/1/012122.
- [2] A. Rahman et al., 'Rainfall Prediction System Using Machine Learning Fusion for Smart Cities', *Sensors*, vol. 22, no. 9, p. 3504, May 2022, doi: 10.3390/s22093504.
- [3] B. Van Den Hurk, F. Doblas-Reyes, G. Balsamo, R. D. Koster, S. I. Seneviratne, and H. Camargo, 'Soil moisture effects on seasonal temperature and precipitation forecast scores in Europe', *Clim. Dyn.*, vol. 38, no. 1–2, pp. 349–362, Jan. 2012, doi: 10.1007/s00382-010-0956-2.
- [4] H. B. Mann, 'Nonparametric Tests Against Trend', *Econometrica*, vol. 13, no. 3, p. 245, Jul. 1945, doi: 10.2307/1907187.
- [5] M. G. Kendall, 'A NEW MEASURE OF RANK CORRELATION', *Biometrika*, vol. 30, no. 1–2, pp. 81–93, Jun. 1938, doi: 10.1093/biomet/30.1-2.81.
- [6] P. K. Sen, 'Estimates of the Regression Coefficient Based on Kendall's Tau', *J. Am. Stat. Assoc.*, vol. 63, no. 324, pp. 1379–1389, Dec. 1968, doi: 10.1080/01621459.1968.10480934.
- [7] I. Krtolica, M. Todorov, S. Prohaska, and M. Stojković, 'Annual and Low-Flow Trends in Serbia', *J. Hydrol. Eng.*, vol. 29, no. 3, p. 05024008, Jun. 2024, doi: 10.1061/JHYEFF.HEENG-6030.
- [8] M. Stojković, S. Prohaska, and J. Plavšić, 'Stochastic structure of annual discharges of large European rivers', *J. Hydrol. Hydromech.*, vol. 63, no. 1, pp. 63–70, Mar. 2015, doi: 10.1515/johh-2015-0009.
- [9] M. Gocic and S. Trajkovic, 'Analysis of changes in meteorological variables using Mann-Kendall and Sen's slope estimator statistical tests in Serbia', *Glob. Planet. Change*, vol. 100, pp. 172–182, Jan. 2013, doi: 10.1016/j.gloplacha.2012.10.014.
- [10] K. H. Hamed and A. Ramachandra Rao, 'A modified Mann-Kendall trend test for autocorrelated data', *J. Hydrol.*, vol. 204, no. 1–4, pp. 182–196, Jan. 1998, doi: 10.1016/S0022-1694(97)00125-X.
- [11] H. X. Do, S. Westra, and M. Leonard, 'A global-scale investigation of trends in annual maximum streamflow', *J. Hydrol.*, vol. 552, pp. 28–43, Sep. 2017, doi: 10.1016/j.jhydrol.2017.06.015.
- [12] Md. Hussain and I. Mahmud, 'pyMannKendall: a python package for non parametric Mann Kendall family of trend tests.', *J. Open Source Softw.*, vol. 4, no. 39, p. 1556, Jul. 2019, doi: 10.21105/joss.01556.



IMPLEMENTATION OF THE DEBUGGING SUPPORT FOR THE LLVM OUTLINING OPTIMIZATION

Vojislav Tomašević^{1*},
[0009-0000-5948-0123]

Đorđe Todorović²,
[0009-0007-6109-8127]

Maja Vukasović¹
[0000-0003-0647-1922]

¹School of Electrical Engineering,
Belgrade, Serbia

²HTEC,
Belgrade, Serbia

Abstract:

Outlining optimization has been used in compilers predominantly to decrease the code size and sometimes even to improve its efficiency. If a code segment is repeated at various places in the code, the compiler can encapsulate it into a function and then it replaces these segments with the function calls. The *LLVM* infrastructure supports the outlining optimization but lacks proper debugging information in such cases and an outlined function cannot be differentiated from some other compiler-generated code at debug time. This paper proposes the complete solution to this problem on three levels of abstraction: *IR* and *Machine IR* code, *DWARF* format, and *LLDB* debugger. It identifies the reason for insufficient precision of the previous debugging information and describes in detail the implementation focused on enhancing these data so that the outlined function can be exactly recognized during debugging. The implementation has been thoroughly tested through regression and custom tests and it was made publicly available. In practice, the enhanced debugging information for outlining has proven to be useful.

Keywords:

Compilers, *LLVM* Infrastructure, Outlining Optimization, Debugging.

INTRODUCTION

In addition to its basic purpose of generating executable code, an important goal of a compiler is to make such code as efficient as possible, regarding both execution speed and memory usage. To this end, numerous optimizations are applied at various places to transform the code into its improved equivalent [1]. One of them is the outlining optimization supported by some compilers. Its main goal is to extract the same, replicated code from different places in the program into a new function and to replace all occurrences of the segment with a corresponding function call. In this way, the entire code size is decreased, and memory consumption is improved.

The compilers need to overcome the high abstraction level and produce efficient code, so they consist of many components. The compilation runs in certain consecutive phases that sometimes partly overlap. Thus, the compilers are quite complex and hard to develop. Because of that, it is very important to have some suitable infrastructure platforms

Correspondence:

Vojislav Tomašević

e-mail:

vojkan99@gmail.com





that make creating the compilers for arbitrary languages and target machines less demanding. They provide ready-made tools and projects that can be used and customized for the development of a proprietary compiler. One of the best-known and widely used platforms is the *LLVM* infrastructure [2]. Besides the tools used in the compilation process, e.g. *Clang* front-end compiler for C and C++ languages, *LLVM* contains rich support for debugging providing a vast amount of data about the program, stored in a convenient form.

LLVM does support the outlining optimization pass, but there is a problem related to information relevant to the debugging process since the outlined code cannot be differentiated from the code generated in some other way. To allow better precision in the interpretation of the code and to enable more efficient debugging, it is necessary to enhance the *LLVM* infrastructure with some additional information that indicates the occurrence and location of the outlined code. This paper proposes a solution to the previous problem by expanding debugging-related information in the context of outlining optimization. The proposal is successfully implemented and thoroughly tested using *LLVM* and *LLDB* testing infrastructures.

The second section of this paper explains the outlining optimization itself, its benefits, and how it is performed. The third section focuses on the solution proposal and describes its implementation on all three levels in detail. The implementation is illustrated with the appropriate code segments. The conclusion summarizes the paper.

2. OUTLINING OPTIMIZATION

Outlining is an optional part of the compilation process that removes a part of the code from a function and replaces it with the call of the new function, which consists of the removed code [3]. The new function is artificially generated by the compiler (more precisely, by its optimizer), and we say that the function is outlined. This process is the opposite of the more frequently employed inlining optimization that embeds the body of a function at its call site, to speed up the program by eliminating the function-call overhead. Since the same function can be inlined at multiple call sites, if not applied selectively, this optimization may cause severe code growth [4].

On the other hand, outlining results in additional calls of the newly created function, which can slow down the execution. However, it pays off by decreasing the code size if the same outlined code is replicated

multiple times. Therefore, outlining optimization is useful predominantly in systems where memory is a critical resource, but which are also fast enough so that the additional function calls don't result in significant performance degradation (e.g., microcontrollers).

Typical applications of outlining are: code refactoring or extraction of the kernel in compilers which transform the source code written in one language to another [5], and shortening of large functions to decrease the compilation time in *JIT* compilers [6]. Another source of potential performance gain can be obtained in cases when a large, frequently called function (hot function) contains some rarely executed regions (cold code) [3]. Outlining of cold code sections from a hot function can have at least three advantages:

1. Removing cold code from a large hot function can make it small enough to apply inlining;
2. Outlining the cold code can improve the cache memory efficiency by preserving the spatial locality of hot code;
3. Outlining the cold code from a hot function can also improve memory bandwidth during instruction fetching, which is important for modern superscalar and *VLIW* architectures.

Outlining optimization in the *LLVM* infrastructure can be carried out on both *IR* and *MIR* code levels. The algorithm of outlining on the *MIR* level consists of the following phases [6]:

1. Identification of the candidates – In this initial phase, all basic blocks [1] in the program are searched for the longest repeated sequence of instructions. This resembles the problem of finding the longest common substring where basic blocks act as strings and instructions act as characters. This problem can be solved by using the suffix tree program representation;
2. Removing unsafe or useless candidates – After the candidates are found, the potential adverse effects of their outlining should be examined (e.g., some instructions like conditional branches cannot be safely extracted from the function). Thus, unsafe candidates are rejected as well as those candidates that do not contribute to the decrease of the code size. The formula for calculating the usefulness of the outlining process is given in [6];
3. Function sharing – After the list of candidates is finalized, code transformation takes place. The new functions are created, and each candidate is replaced with the corresponding calls.



3. IMPLEMENTATION

Although the *LLVM* infrastructure supports the outlining optimization, its implementation is not complete enough as far as the debugging information is concerned. As for the debugging data, *LLVM* treats the new functions generated in case of outlining the same way as any other compiler-generated code. This shortcoming motivated us to propose a solution within the *LLVM* that makes a distinction between outlined code and other compiler-generated code [7]. This solution required some modifications on three levels of abstraction: *IR* and *MIR* code, *DWARF* format, and *LLDB* debugger. This proposal covers a wide range of the *LLVM* debug information that is useful for both the user and programmer to see whether the outlining is applied somewhere in the code. Support for these three levels is presented in the following text. The entire implementation is publicly available [8].

3.1. THE *IR* AND *MIR* CODE LEVELS

IR (*intermediate representation*) code used in *LLVM* is generated by the front-end and the middle-end (i.e., the optimizer) of the compiler [9]. Its advantages are the architectural independence and separation between the compiler front-end and back-end, which enables the different implementations of these two parts to connect and work together. This kind of intermediate code is very suitable for optimizations. *MIR* (*machine-specific*

intermediate representation) is also a type of intermediate code that is used in *LLVM* in the compiler back-end. It is generated after the instruction selection phase of the compilation. *MIR* code is also very convenient for target architecture-aware optimizations that are performed in the back-end. When debugging information is enabled, *IR* code keeps it in the form of the *LLVM* metadata. Since *MIR* is an extension of *IR* and each *MIR* module contains a corresponding *IR* module, *MIR* refers to the same metadata with debug information from the contained *IR* module. All metadata in the *LLVM* infrastructure represents class objects derived from the *llvm::Metadata* base class [10].

Outlining optimization in the *LLVM* infrastructure is implemented on both *IR* and *MIR* levels at the module level in their *runOnModule* functions. Besides the generation of an outlined function, its corresponding metadata node *DISubprogram* is also created by calling *createFunction* of the *DIBuilder* class. For this node, there is the *DINode::DIFlags::FlagArtificial* flag that is set to indicate that the function is artificially created by the compiler. Listing 1 shows a part of the *IR* file for the sample program after outlining two functions *outlined_ir_func_0* and *outlined_ir_func_1*. The *DISubprogram* metadata nodes for them contain *DIFlagArtificial* indicator which informs that these functions are artificially generated by the compiler. However, it is not clear that they are outlined.

```
; Function Attrs: nounwind uwtable
define dso_local i32 @main() #0 !dbg !7 {
entry:
    %x = alloca i32, align 4
    %y = alloca i32, align 4
    call void @outlined_ir_func_1(i32* %x), !dbg !14
    call void @outlined_ir_func_1(i32* %y), !dbg !14
    call void @outlined_ir_func_0(i32* %x, i32* %y), !dbg !14
    call void @outlined_ir_func_0(i32* %y, i32* %x), !dbg !14
    ret i32 0
}

!15 = distinct !DISubprogram(name: "outlined_ir_func_0", linkageName: "outlined_ir_func_0",
scope: !1, file: !1, type: !16, flags: DIFlagArtificial, spFlags: DISPFlagDefinition |
DISPFlagOptimized, unit: !0, retainedNodes: !17)
!16 = !DISubroutineType(types: !17)
!17 = !{}
!18 = !DILocation(line: 0, scope: !15)
!19 = distinct !DISubprogram(name: "outlined_ir_func_1", linkageName: "outlined_ir_func_1",
scope: !1, file: !1, type: !16, flags: DIFlagArtificial, spFlags: DISPFlagDefinition |
DISPFlagOptimized, unit: !0, retainedNodes: !17)
```

Listing 1. A part of the resulting *IR* file after the outlining is applied to a sample program [7]



The solution to this problem is found in the addition of a new flag named *Outlined* in the *DIFlags enum* statement. This *enum* statement contains all flags for some metadata and it is a part of the *DINode* class also inherited by the *DISubprogram* class. The role of the new flag is to reflect the information that the corresponding *DISubprogram* is an outlined function. Like all other flags, the *Outlined* flag is declared in *llvm/include/llvm/IR/DebugInfoFlags.def* file, whose content is included in the *enum* statement. The flag was then added by creating an appropriate call of the *HANDLE_DI_FLAG* macro, which has previously been defined in the *DIFlags enum* statement. After that, the definition of the *Largest* flag was also updated accordingly. These code modifications are shown in Figure 1.

After adding the new flag, its use on both *IR* and *MIR* levels was implemented. When an outlined function and its corresponding *DISubprogram* object are created, instead of a more general *DINode::DIFlags::FlagArtificial* indicator which tells that the function does not exist in the source code, the new *DINode::DIFlags::FlagOutlined* flag is set. It provides more precise information about the nature of the compiler-generated function, as illustrated in Figure 2. Modification on the *MIR* level was performed in a quite similar manner.

Now, after the outlining is applied, debugging information in the previous example is shown in Listing 2.

3.2. DWARF FORMAT LEVEL

The *DWARF* format is one of the most frequently used formats for debugging information [11]. According to *DWARF*, debug information is represented as a tree-like structure. Hence, it consists of the *DIE* nodes connected by basic parent-children relationships. The characteristics of program entities represented by the *DIE* objects are described by a versatile set of attributes of different types (can even be a reference to some other *DIE* node).

One of the *DIE* objects is also the *DW_TAG_subprogram* object which represents the function. Among other attributes, it contains the *DW_AT_artificial* attribute whose purpose is to denote the compiler-generated construct that does not exist in the source code. Again, this information is not precise enough in the case of outlined functions. Therefore, the outlined code needs to be distinguished from some other artificially generated code.

```
3  llvm/include/llvm/IR/DebugInfoFlags.def
@@ -58,6 +58,7 @@ HANDLE_DI_FLAG((1 << 26), NonTrivial)
58  HANDLE_DI_FLAG((1 << 27), BigEndian)
59  HANDLE_DI_FLAG((1 << 28), LittleEndian)
60  HANDLE_DI_FLAG((1 << 29), AllCallsDescribed)
61  + HANDLE_DI_FLAG((1 << 30), Outlined)
62
63  // To avoid needing a dedicated value for IndirectVirtualBase, we use
64  // the bitwise or of Virtual and FwdDecl, which does not otherwise
@@ -67,7 +68,7 @@ HANDLE_DI_FLAG((1 << 2) | (1 << 5), IndirectVirtualBase)
67  #ifdef DI_FLAG_LARGEST_NEEDED
68  // intended to be used with ADT/BitmaskEnum.h
69  // NOTE: always must be equal to largest flag, check this when adding new flag
70  - HANDLE_DI_FLAG((1 << 29), Largest)
71  + HANDLE_DI_FLAG((1 << 30), Largest)
72  #undef DI_FLAG_LARGEST_NEEDED
73  #endif
```

Figure 1. Introducing the new *Outlined* flag [8]

```
2  llvm/lib/Transforms/IPO/IROutliner.cpp
@@ -624,7 +624,7 @@ Function *IROutliner::createFunction(Module &M, OutlinableGroup &Group,
624  0 /* Line 0 is reserved for compiler-generated code. */,
625  DB.createSubroutineType(DB.getOrCreateTypeArray(None)), /* void type */
626  0, /* Line 0 is reserved for compiler-generated code. */
627  - DINode::DIFlags::FlagArtificial /* Compiler-generated code. */,
628  + DINode::DIFlags::FlagOutlined /* Compiler-generated outlined code. */,
629  /* Outlined code is optimized code by definition. */
630  DISubprogram::SPFlagDefinition | DISubprogram::SPFlagOptimized);
```

Figure 2. Using the *Outlined* flag on the *IR* level [8]



```
!15 = distinct !DISubprogram(name: "outlined_ir_func_0", linkageName: "outlined_ir_func_0",
scope: !1, file: !1, type: !16, flags: DIFlagOutlined, spFlags: DISPFlagDefinition |
DISPFlagOptimized, unit: !0, retainedNodes: !17)
!16 = !DISubroutineType(types: !17)
!17 = !{}
!18 = !DILocation(line: 0, scope: !15)
!19 = distinct !DISubprogram(name: "outlined_ir_func_1", linkageName: "outlined_ir_func_1",
scope: !1, file: !1, type: !16, flags: DIFlagOutlined, spFlags: DISPFlagDefinition |
DISPFlagOptimized, unit: !0, retainedNodes: !17)
```

Listing 2. *DISubprogram* metadata of the outlined functions from Listing 1 after *Outlined* flag is introduced [7]

```
1  llvm/include/llvm/BinaryFormat/Dwarf.def
599 599 HANDLE_DW_AT(0x3e03, LLVM_tag_offset, 0, LLVM)
600 600 // The missing numbers here are reserved for ptrauth support.
601 601 HANDLE_DW_AT(0x3e07, LLVM_apinotes, 0, APPLE)
602 602 + HANDLE_DW_AT(0x3e08, LLVM_outlined, 0, LLVM)
603 603
604 604 // Apple extensions.
605 605
```

Figure 3. Declaration of the *DW_AT_LLVM_outlined* DWARF attribute [8]

```
3  llvm/lib/CodeGen/AsmPrinter/DwarfUnit.cpp
1313 1313 void DwarfUnit::applySubprogramAttributes(const DISubprogram *SP, DIE &SPDie,
1314 1314 addThrowTypes(SPDie, SP->getThrowTypes());
1315 1315
1316 1316 + if (SP->isOutlined())
1317 1317 + addFlag(SPDie, dwarf::DW_AT_LLVM_outlined);
1318 1318 +
1319 1319 if (SP->isArtificial())
1320 1320 addFlag(SPDie, dwarf::DW_AT_artificial);
1321 1321
```

Figure 4. Adding the *DW_AT_LLVM_outlined* attribute to the *DIE* object of the function [8]

Translation of a program in *LLVM* acts as a pipeline. It enables the propagation of information generated in an earlier phase of compilation to the later stages. This property was exploited in augmenting the debugging information regarding outlining optimization in the *DWARF* format.

Since the new, *Outlined* indicator is added on the *IR* level, it is propagated through the compiler up to the place where the support for building the debugging information in *DWARF* format is implemented as an indication that some function is outlined. It is achieved by introducing the *isOutlined* function in the *DISubprogram* class which checks whether the *Outlined* indicator for a given object is set. If the call of the *isOutlined* function

returns a *true* value, the attribute *DW_AT_LLVM_outlined* is added to the corresponding *DIE* object representing the outlined function. This attribute is introduced in the *DWARF* format by adding another call of the *HANDLE_DW_AT* macro with the *LLVM_outlined* argument. It was done in the part reserved for the extension of the format for the *LLVM* project needs, in the file *llvm/include/llvm/BinaryFormat/Dwarf.def* file as shown in Figure 3.

The addition of this attribute in the *DIE* object related to the *DISubprogram* metadata is shown in Figure 4. Listing 3 shows the contents of the *DW_TAG_subprogram* objects of the two outlined functions from Listing 1 after the proposed modifications are implemented.



```
0x0000005b:    DW_TAG_subprogram
                DW_AT_low_pc      (0x000000000000004e)
                DW_AT_high_pc     (0x0000000000000052)
                DW_AT_frame_base  (DW_OP_reg7 RSP)
                DW_AT_linkage_name ("outlined_ir_func_1")
                DW_AT_name        ("outlined_ir_func_1")
                DW_AT_LLVM_outlined (true)
                DW_AT_external    (true)

0x00000065:    DW_TAG_subprogram
                DW_AT_low_pc      (0x000000000000003c)
                DW_AT_high_pc     (0x000000000000004e)
                DW_AT_frame_base  (DW_OP_reg7 RSP)
                DW_AT_linkage_name ("outlined_ir_func_0")
                DW_AT_name        ("outlined_ir_func_0")
                DW_AT_LLVM_outlined (true)
                DW_AT_external    (true)
```

Listing 3. New display of the *DW_TAG_subprogram* objects for two outlined functions [7]

3.3. LLDB DEBUGGER LEVEL

LLDB debugger is also developed as a part of the *LLVM* project [12]. Like every other debugging tool, it accepts executable code as input and then executes it using the debugging information, with the provided parameters. During the debugging process, in some situations, *LLDB* prints messages to the user that help him understand the program being executed, as well as the modifications of the program resulting from applied optimizations. Enhancing this information regarding outlining optimization in the *LLDB* debugger is the main goal at this level.

The problem is quite like to the problems we have faced on the *IR/MIR* and *DWARF* levels. If a function is outlined, the *LLDB* tool recognizes it as artificially generated by the compiler, without any further explanation. Namely, during the execution of an outlined function, the *LLDB* debugger prints the following message „*Note: this address is compiler-generated code that has no source code associated with it.*“ This is a generic message which is printed each time some compiler-generated code is encountered. If the name of the function with the compiler-generated code is known, it is also given in the previous message as a parameter. Whether some code is compiler-generated or not is checked by the *GetStatus* method of the *StackFrame* class, which prints the description of the stack frame and/or source (or assembly) code context for this frame. This method examines whether there is a source code line that corresponds to the current address during debugging. If the *line* field of the *line_entry* object of the *LineEntry* class that represents a field of the *m_sc* object of the *SymbolContext* class is equal to zero, it is an indication of an artificially generated code.

To solve the previous ambiguity, it is designed that the *LLDB* debugger prints a new, custom message when an outlined function is encountered. Unlike previous messages about compiler-generated code that appear only once, the new message is printed after each instruction of an outlined function is executed, provided that step-by-step debugging mode is active.

Implementation of this *LLDB* debugger extension is based on the *Function* class since it represents a connection between the *DWARF* format and the *LLDB* as far as the function debugging is concerned. It contains information on whether the function represented by the object of this class is outlined. For this purpose, a new *m_outlined* field is added as well as the corresponding parameter of the class constructor for the initialization of the new field (the default value is *false*). Also, this class is extended with the *IsOutlined* method that returns the value of the *m_outlined* field.

Finally, it was necessary to implement the support for setting the new field in the following way. By parsing the die *DWARFDIE* object, the *ParseFunctionFromDWARF* method of the *DWARFASTParserClang* class creates the object of the *Function* class and returns the pointer to it. This *die* must have the *DW_TAG_subprogram* tag, so that it represents a function. During the parsing of the object, a *for* loop that traverses all *DWARF* attributes of the object is added and, if the new *DW_AT_LLVM_outlined* attribute is found, it keeps the record that the function is outlined by setting a local variable to *true*. This variable is later forwarded to the *Function* class constructor as a parameter that initializes the *m_outlined* field.



In this way, the information of the applied outlining optimization is propagated from the *DWARF* format to the *LLDB* tool. Previous implementation details are presented in Figure 5.

After the object of the *Function* class is created, it can be used to detect whether some function is outlined or not. This possibility is exploited in the *GetStatus* method of the *StackFrame* class. Besides reporting two previous messages about the existence of the artificially generated code, the new message „*Note: this function is outlined.*“ is displayed in case the *IsOutlined* method for the current function returns value *true*. This change is illustrated in Listing 4.

4. CONCLUSION

The debugging process is crucial for software testing and its efficiency essentially depends on the completeness and the precision of the debugging information. In the *LLVM* project, debug data cannot recognize the situation when a function is compiler-generated by outlining. This paper proposes a solution to this problem by enhancing the debugging information and its appropriate handling. The proposal is carefully implemented on the three levels of abstraction: the *IR* and *MIR* code, *DWARF* format, and *LLDB* debugger. Thorough regression testing was carried out to verify the correctness of the solution. It required enhancing the existing tests and writing new ones.

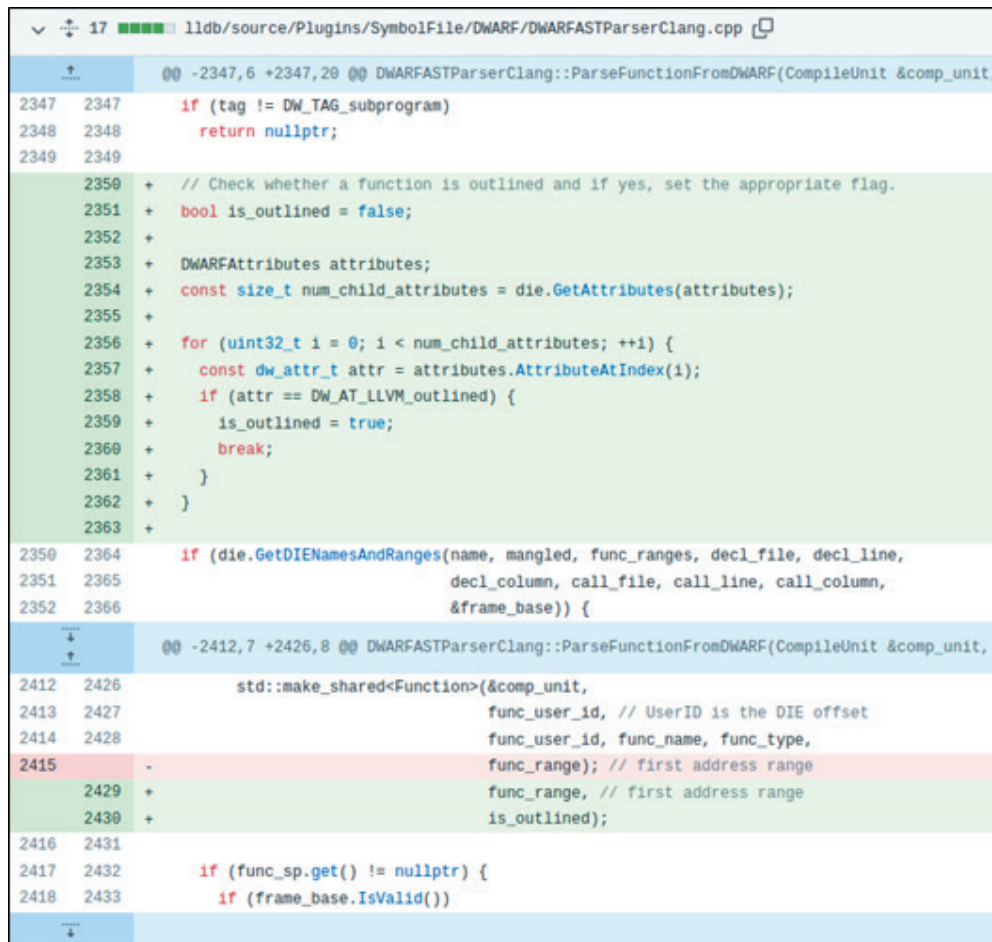


Figure 5. Propagation of information about outlining from *DWARF* to *LLDB* [8]

```
if (m_sc.function && m_sc.function->IsOutlined()) {
    strm.Printf("Note: this function is outlined.");
    strm.EOL();
}
```

Listing 4. The new message for an outlined function in the *GetStatus* method of the *StackFrame* class [7]



The implementation of the support for enhancing the debugging information in the context of the outlining optimization does not require the writing of a large amount of code, as can be seen from the proposed solution. Instead, it required a very demanding analysis of a complex project like *LLVM* and inserting many small changes at different places in the code to encompass all levels of abstraction in the *LLVM* project followed by exhaustive testing. Hence, the main part of the task is of the research type, while the implementation is a less demanding part. It is a typical pattern in the *LLVM* infrastructure when the modification or addition of a relatively small amount of code can provide a significant effect. There is also an idea for further improvement of user experience during debugging in *LLDB*. It can be achieved by providing support for the reconstruction of local variables in the outlined functions.

REFERENCES

- [1] M. L. Scott, *Programming Language Pragmatics*, Morgan Kaufman, 2006.
- [2] "The LLVM Compiler Infrastructure," [Online]. Available: <https://llvm.org/>. [Accessed 27 March 2025].
- [3] P. Zhao and J. N. Amaral, "Function Outlining," Dept. of Computing Sciences, Univ. of Alberta, Edmonton, Canada, 2010.
- [4] M. Vukasović and A. Prokopec, "Exploiting Partially Context-sensitive Profiles to Improve Performance of Hot Code," *ACM Transactions on Programming Languages and Systems*, vol. 45, no. 4, pp. 1-64, 2023.
- [5] C. Liao, D. J. Quinlan, R. Vuduc and T. Panas, "Effective Source-to-Source Outlining to Support Whole Program Empirical Optimization," in *International Workshop on Languages and Compilers for Parallel Computing*, Newark, DE, United States, 2009.
- [6] "Reducing code size with LLVM Machine Outliner on 32-bit Arm targets," [Online]. Available: <https://www.linaro.org/blog/reducing-code-size-with-llvm-machine-outliner-on-32-bit-arm-targets/>. [Accessed 30 August 2022].
- [7] V. M. Tomašević, "Unapređenje infrastrukture LLVM dodavanjem informacija za otklanjanje grešaka prilikom autlajning optimizacije," School of Electrical Engineering, University of Belgrade, 2022.
- [8] "[Outliner] Add debug-info support in IR, DWARF and LLDB," [Online]. Available: <https://github.com/llvm/llvm-project/commit/80e1c808dd121595f7124917dd7ef22bb0da5fa7?diff=unified>. [Accessed 25 March 2025].
- [9] "Opt – LLVM optimizer," [Online]. Available: <https://llvm.org/docs/CommandGuide/opt.html>. [Accessed 20 March 2025].
- [10] "llvm::Metadata Class Reference," [Online]. Available: https://llvm.org/doxygen/classllvm_1_1Metadata.html. [Accessed 20 March 2025].
- [11] "DWARF Debugging Information Format Version 5," [Online]. Available: <https://dwarfstd.org/doc/DWARF5.pdf>. [Accessed 17 March 2025].
- [12] "The LLDB Debugger," [Online]. Available: <https://lldb.llvm.org/>. [Accessed 23 March 2025].



AI AND CYBERSECURITY IN E-COMMERCE: HOW AI INTEGRATION AFFECTS DATA PROTECTION

Emilija Jovanović^{1*},
[0009-0003-7007-6537]

Mladen Veinović¹,
[0000-0001-6136-1895]

Miloš Jovanović²
[0009-0008-9032-8195]

¹Singidunum University,
Belgrade, Serbia

²Faculty of Mechanical and
Civil Engineering in Kraljevo,
University of Kragujevac,
Kragujevac, Serbia

Abstract:

Every day, artificial intelligence sets significant innovative changes in all fields. Technological innovations introduced by AI have a positive effect on the protection of users in cyber security. Using data analysis, AI can quickly recognize threats and identify signs for early intervention. AI has a crucial role in protecting user accounts, detecting illegal transactions and preventing financial fraud in the sphere of e-commerce.

Despite all the benefits, the use of artificial intelligence also brings problems, such as false positive results resulting from errors of the machines themselves, and high implementation costs. The fact that must not be ignored is that the human factor is still crucial despite everything.

The paper focuses on exploring the application of artificial intelligence in e-commerce cyber security, particularly in the detection of suspicious transactions. An experimental model of Random Forest machine learning algorithms was used to evaluate suspicious transactions in e-commerce. The paper highlights the advantages and challenges of applying AI, emphasizing the importance of integrating intelligent systems and human analysis, which achieves effective and safe protection of the digital space.

Keywords:

E-commerce, AI, Cybersecurity, Data Protection.

INTRODUCTION

The development of electronic commerce in the first decades of the 21st century brought numerous benefits to both consumers and companies, providing faster and easier shopping, significant improvement of global business and automation of sales processes. However, the development of electronic commerce has been accompanied by an increase in cyberattacks.

Traditional data protection methods, such as encryption and multi-factor authentication, have not been sufficient in recent years to fully protect user data from attacks using automated tools and artificial intelligence. For these reasons, AI is increasingly being used as an important technology in cyber security, enabling faster and easier detection of threats, proposing security measures according to potential threats and improving data protection in real time.

Correspondence:

Emilija Jovanović

e-mail:

ejovanovicrs@gmail.com





This paper explores how the application of artificial intelligence can improve cyber security in e-commerce, with a special focus on protecting user data and preventing potential cyberattacks. By analyzing modern AI technologies, we concluded that machine learning algorithms, anomaly detection and user behaviour monitoring identify early threats, automate the response to potential threats and improve security standards.

Also, the paper will address the topic of challenges and limitations of the application of artificial intelligence in cyber security, with special reference to ethical dilemmas, regulatory obstacles and multiple risks. A special emphasis will be on adapting AI technologies to current security frameworks, such as NIST and ISO standards, to ensure more adequate data protection and strengthen user confidence in the electronic trading platforms themselves.

The main task of this paper is to show the advantages and disadvantages of the application of artificial intelligence, as well as to give an insight into possible challenges during the implementation of artificial intelligence in cybersecurity. [1]

2. CYBER-SECURITY RISKS IN E-COMMERCE

2.1. TRENDS IN CYBER-ATTACKS

The digitalization of business and the increasing number of online transactions have contributed to the increase in the number of cyberattacks. Research has shown that a large number of cyberattacks occurred worldwide in one day, with the majority of them targeting e-commerce platforms, banking systems and financial sectors. Nowadays, innovative methods are used to steal data and financial assets, and the most vulnerable sectors are healthcare, logistics, retail and financial institutions. Companies that do not use advanced data protection methods are in great danger, which is why more and more investments are being made in advanced protection systems to reduce the risk of potential cyberattacks. [2]

2.2. TYPES OF CYBERATTACKS

Nowadays, protecting online stores from cyberattacks is an increasing challenge. Hackers skillfully use various modern methods to compromise the security of data and financial activities.

Phishing is considered one of the most common cyber-attacks. In order to obtain sensitive data, such as passwords, credit card numbers or personal information,

attackers use fake e-mail addresses, messages or websites. Mostly, these attacks rely on social momentum, using well-known brands or institutions that have gained user trust and thus encourage them to take action. Innovative forms of phishing, such as spear phishing, use personalized messages to target individuals or organizations in order to carry out a successful attack.

A serious threat is DDoS (Distributed Denial-of-Service) attacks, because the servers are occupied by a large number of requests and thus prevent legitimate users from accessing it. These attacks result in a major financial crash, reduce trust and damage the company's reputation. To carry out this type of attack, botnet networks - networks of compromised computers - are used, which generate a large amount of content and destabilize the system.

Malware, malicious software, violates the protection of user data, redirects transactions or takes over system management. Malicious programs such as viruses, trojans, ransomware and spyware are used to carry out such an attack. They are mostly spread through infected e-mails, fake advertisements or downloaded files. Sensitive information such as phone numbers, credit card numbers and passwords can be compromised through the use of malware, directly compromising the security of both buyers and sellers. [3]

2.3. REGULATORY FRAMEWORK: OVERVIEW OF NIST AND ISO DATA PROTECTION STANDARDS

To secure data, organizations apply a regulatory framework for data protection that includes various standards and guidelines. The two most important sets of standards in the field of data protection are NIST (National Institute of Standards and Technology) and ISO (International Organization for Standardization) standards. These standards serve to help when implementing systems for effective data protection as well as systems that provide data security.

NIST is an American agency that develops standards, guidelines, and methods in various fields. These standards are mainly applied to federal information systems in the United States but are also applicable in a wider range of sectors. ISO is an international organization that develops global standards for various industries. The ISO standard is more globally accepted regarding security, but NIST is also used more in technical implementation.



3. THE IMPACT OF ARTIFICIAL INTELLIGENCE IN E-COMMERCE CYBER SECURITY

3.1. HOW DOES AI HELP DETECT AND PREVENT ATTACKS?

AI (artificial intelligence) is an integral part of detecting and preventing attacks in cyber security today. Artificial intelligence and machine learning are winning techniques as they enable systems to detect threats in real time, respond and also improve security. Traditional security systems use static definitions and signals of attack types, which are often effective against new but standard threats. AI, in this way, enables security tools to learn from huge data sets and identify patterns that indicate threats, long before attacks are actually known.

Real time analysis of large amounts of data is one of the key strengths of artificial intelligence in cybersecurity. In this way, faster detection of threats is possible due to the recognition of unusual activities, such as unusual access to data, variations in network traffic or unusual changes in user behaviour. The ability to stop attacks in their initial stages can greatly reduce the damage that attackers can inflict.

Automated analysis and classification of attacks are enabled by artificial intelligence. This means that the response to potential threats can sometimes be accelerated. In this case, artificial intelligence will monitor the network. When it detects an attack, it will act immediately, however necessary to meet those challenges. This can mean blocking access to the system, separating infected devices from uninfected parts, and redirecting suspicious traffic elsewhere.

Data encryption also benefits from the use of artificial intelligence. The integration of advanced AI algorithms into encryption helps to detect and prevent potential vulnerabilities in data protection systems, so that data is truly protected and attackers cannot find loopholes in existing protection methods.

3.2. KEY AI METHODS

3.2.1. Anomaly detection

Anomaly detection in security and protection is where artificial intelligence is particularly useful. By recognizing unusual patterns in the data, it is possible to detect potential threats. Traditional security systems are based on pre-defined correct and known threats, but this is not enough because different attack methods are evolving very quickly.

AI and machine learning can indirectly analyze user behaviour, and identify suspicious activities even through abnormal changes in activity or unusual transactions.

One of how AI contributes to the detection of anomalies is through User Behaviour Analytics (UBA), that is, the analysis of user behaviour. For example, machine learning algorithms can track a user's normal ways of using systems and applications, noting their usual activity patterns. The system can monitor login time, IP address, location and device from which the user connects, as well as typical user operations: creating new files or modifying old ones. When there is a sudden change in this pattern, that is, if any deviation from these habits is observed - for example, when a user logs in from an unknown place at an unusual time one day or suddenly tries to download a very large file - the AI can generate an alert, temporarily lock the account and require additional authentication as necessary. [4]

3.2.2. Behavioural authentication

Biometric authentication is all about analysing and measuring deviations in human behaviour to verify a specific individual. While conventional biometrics are based on something that a user has (fingerprint, iris scan, or facial recognition), behavioural biometrics analyze the way that a user accesses a device. Each individual demonstrates a unique set of behaviour patterns that can be quantified, making them an ideal security instrument. [5]

3.2.2.1. Dynamics of typing on the keyboard

The typing style of a person can identify them. The AI assesses elements including typing rhythm, keystroke speed and the time between pressing and releasing a key (known as "time of flight"). Every user will have a unique text input pattern and any deviation from the normal one could indicate a breach.

3.2.2.2. Mouse movement and touchpad interaction

Mouse movements also depend on the way a user moves it, the speed of the movement, how it moves — in a moon or more smooth, even pace — the number of clicks and their precision can prove to be unique elements of your identity. AI can identify these patterns and look for deviations from them to find a possible hacking attempt.



3.2.2.3. *Gait and movement patterns*

Smartphones and wearables utilize sensors to determine how a user moves, regarding stride length, speed of walking, and body posture. These patterns are fairly stable, and deviations from them could be a signal of fraudulent behaviour or that a particular device has been compromised.

3.2.2.4. *Characteristics of speech*

Voice recognition system identifies the user using tonality, intonation, accent and rhythm of speech. This process is commonly used in voice-activated systems, like assistants or banking services, where it will allow the system to recognize its user without additional passwords.

3.2.2.5. *Touch and swipe patterns*

On smartphones and tablets, the way a user touches the screen — pressure strength, swiping speed and scrolling patterns — can provide an added layer of authentication. And that's the reason why it's very hard to fake this data as each user has its own way of interacting with the screen and it makes the system very secure.

By combining these methods, behavioural authentication drastically increases the level of protection and enables early detection of suspicious activities and suppression of unauthorized access. [6]

4. EXPERIMENTAL TESTING OF AI MODELS IN THE DETECTION OF SUSPICIOUS TRANSACTIONS IN E-COMMERCE

The experiment was carried out to examine the possibility of applying artificial intelligence in the detection of suspicious transactions in e-commerce, all based on the analysis of significant parameters such as the user's location, device, purchase time and transaction amount.

To experiment, a machine learning model was used that recognizes transactions whether they are suspicious or not. The data is generated based on real scenarios in e-commerce, where different patterns are simulated. The model analyzes four key factors:

1. **User location** - If the user has made several transactions from one location, and a new transaction comes from another (e.g. from another country), this can be an indicator of potential fraud.
2. **The device from which the transaction was made** - A sudden change in the device from which the transaction was mainly made can signal an abused account.
3. **Time of transaction** – Transactions that occur at unusual times of day, such as late at night, can potentially be risky.
4. **Purchase Amount** – Unexpected and significant changes in transaction value may indicate fraudulent activity.

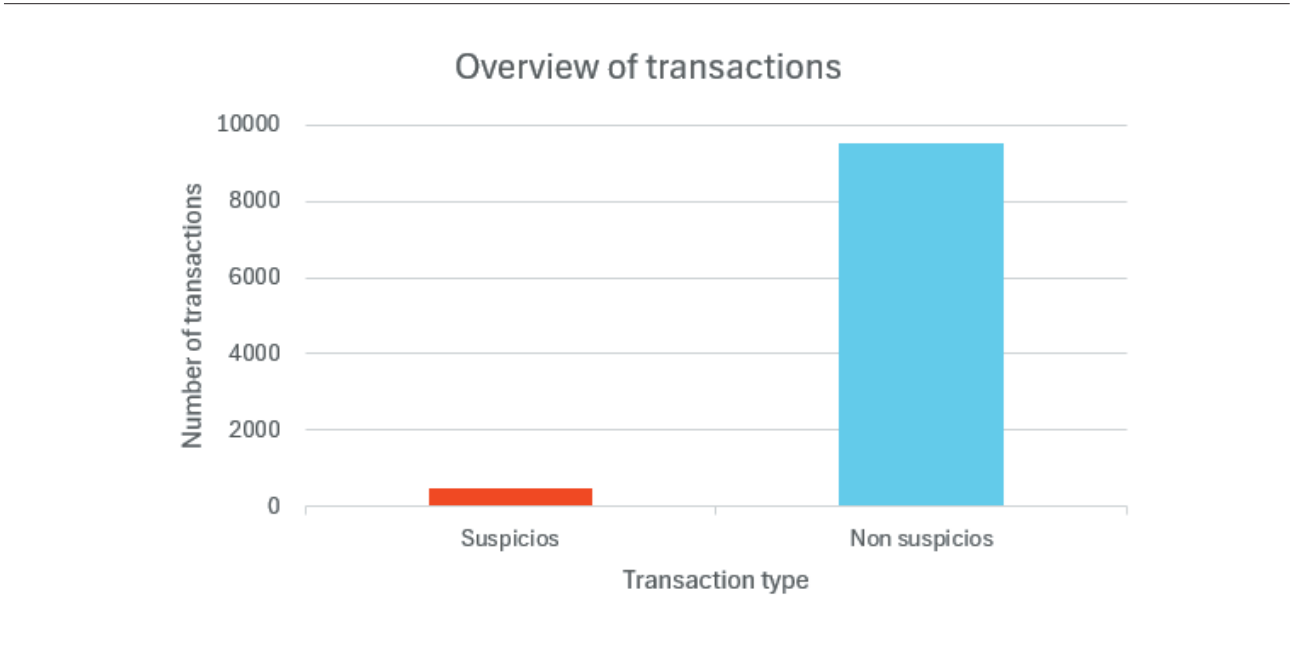


Figure 1. Overview of transaction



To recognize transactions as suspicious or not, the Random Forest algorithm was used, which is known for its ability to recognize complex patterns in data. The model was implemented in the Python programming language using the scikit-learn libraries for machine learning and pandas for data processing. A synthetic dataset of 10,000 generated transactions was used to train the model, of which about 5% were identified as potentially suspicious. The dataset was constructed to simulate real-world e-commerce conditions, including deliberately inserted anomalies so that the model could recognize illegality.

After training, the model was evaluated for accuracy and responsiveness. Figure 1 shows the result visualized through a chart of transactions. This approach enables the automated recognition of potential fraud in real time, contributing to greater platform security and reducing financial risks associated with unauthorized transactions.

5. LIMITATIONS AND CHALLENGES OF IMPLEMENTING AI IN CYBERSECURITY

AI has much to offer cybersecurity, but implementing it is far from simple. Beyond technical limitations, organizations may face ethical, legal and operational challenges that can stall the effective integration of AI in data protection.

5.1. RESULTS FOR FALSE POSITIVES AND FALSE NEGATIVES

AI systems in cyber security also face the challenge of threat detection accuracy. If AI raises too many false alarms (false positives), security teams will begin to disregard notifications. If the AI fails to detect certain threats (false negatives), attackers can exploit the system with impunity.

5.2. LIMITATIONS IN MODEL TRAINING

An AI system's efficacy directly correlates with the quality of the data it is fed. AI can make mistakes if the data provided to it is not suitable or is incomplete, outdated or biased. Furthermore, cyberattacks are continually changing, so models have to be updated and adjusted to new threats in real time.

5.3. AI AS A TARGET OF ATTACK

Attackers are increasingly attempting to hack AI systems, to deceive or compromise them. One of the techniques is "data poisoning", where wrong data is introduced into the system, which leads the AI to incorrect conclusions.

5.4. HIGH COST

Building AI solutions for cybersecurity needs a massive investment in infrastructure, employee training, and continual model development.

5.5. PRIVACY AND REGULATION

AI-based cybersecurity systems typically deal with a large volume of personal and sensitive data, which can, in turn, create privacy and regulatory concerns with laws such as GDPR. A corresponding legal framework is necessary to prevent possible abuses or over-collection of data.

5.6. OVERVIEW OF LIMITATIONS AND CHALLENGES

However, while AI will continue to drive a major evolution in the cyber security landscape, human skills cannot be fully replaced as analysis, prioritisation and response to complex threats will always require human decision-making. AI can automatically identify suspicious activity, but human expertise is needed to analyze data, identify new types of attacks and make decisions about how to protect systems.

The downside of AI systems is that they are only as good as the data they are trained on. When the AI encounters threats that were unknown to it during training, it is the human category that can identify the problem and change the security strategy to mitigate this threat.

The human factor proves central to ethical and legal issues in cyber security, as well as technical aspects. Data privacy, regulation and ethical challenges are deeply human decisions that cannot be handed over to algorithms.

To that end, the best way to achieve cybersecurity is not with AI or humans alone, but with both. Coming together, automated analysis with human expertise will enable faster, more accurate and more secure data protection in an increasingly complex digital environment. [7]



6. CONCLUSION

This paper shows the importance of artificial intelligence in the field of cyber security. All benefits such as anomaly detection, and behavioural authentication, contribute to a significant improvement in the security of digital systems. Specifically in e-commerce, AI has the ability to analyze massive amounts of data in real time, revealing the potential for an attack.

However, despite all the benefits, the implementation of AI brings security challenges. False results can destroy the systems of certain corporations as well as the artificial intelligence system itself. Also, high costs and legal restrictions hinder widespread implementation, especially in smaller organizations. The lack of experts further complicates the development and maintenance of advanced AI solutions.

A winning combination is artificial intelligence and existing security regulations and practices. In this way, organizations can significantly improve data protection and system resilience. The future of cyber security is reflected in intelligent, adaptive systems that will react promptly and also predict potential attacks. The continued development of artificial intelligence, along with ethical and legal oversight, will enable a safer digital environment.

REFERENCES

- [1] S. S. G. S. M. Ashish Juneja, "Cyber Security and Digital Economy: Opportunities, Growth and Challenges," *Journal of Technology Innovations and Energy*, 2024.
- [2] D. R. Gupta, "Cybersecurity Threats in E-Commerce: Trends and Mitigation Strategies," *Journal of Advanced Management Studies*, vol. 1, no. 3, pp. 1-10, 2024.
- [3] A. Bendovschi, "Cyber-Attacks – Trends, Patterns and Security Countermeasures," *7th International Conference On Financial Criminology*, pp. 24-31, 2015.
- [4] A. J.-A. A. M. N. a. A. M.-N. William Villegas-Ch, "Integrating Explainable Artificial Intelligence in Anomaly Detection for Threat Management in E-Commerce Platforms," *IEEE Access*, vol. 6, 2016.
- [5] W. Liang and F. Hamzah, "Behavioral Biometrics and AI for Cloud User Authentication," 2025.
- [6] N. Aisyah, R. Hidayat, S. Zulaikha, A. Rizki, Z. B. Yusof and D. P. a. F. Ismail, "E-Commerce Authentication Security with AI: Advanced Biometric and Behavioral Recognition for Secure Access Control," *QuestSquare*, vol. 7, pp. 55-67, 2021.
- [7] M. F. Ansari, B. Dash and P. S. a. N. Yathiraju, "The Impact and Limitations of Artificial Intelligence in Cybersecurity: A Literature Review," *International Journal of Advanced Research in Computer and Communication Engineering*, 2022.



ADVANCED TECHNOLOGIES AND APPLICATIONS SESSION





HYPERSONIC TARGET DETECTION USING EOIR-EMULATED SYNTHETIC DATA: A CASE STUDY OF X-43A

Slobodan Pajić^{1*},
[0000-0002-6883-8550]

Gerrit Viljoen¹,
[0009-0007-3825-8444]

Marko Tanasković²
[0000-0003-3592-0598]

¹Halcon, EDGE group,
Abu Dhabi, UAE

²Singidunum University,
Belgrade, Serbia

Abstract:

Detecting hypersonic vehicles is challenging due to their high flight velocity, unpredictable flight paths, and distinct thermal signatures. This study used the ANSYS Systems Tool Kit (STK) to generate high-fidelity synthetic datasets. These datasets were designed to simulate Electro-Optic Infrared (EOIR) scenarios tailored specifically for hypersonic targets. Inspired by the X-43A hypersonic flight tests, we created a simplified environment without terrain modeling to balance computational efficiency and physics-based simulation. Our approach tackles the critical lack of real-world data by offering a scalable and realistic foundation for advancing Automatic Target Recognition (ATR) systems. Our simulations demonstrated that a Modified HALO-II sensor achieved an 18% improvement in detection rates compared to the standard HALO-II sensor, even under challenging thermal conditions. We explain the experimental setup and ATR dataset generation. Future work will focus on integrating synthetic with experimental datasets to help AI-driven detection systems address the evolving challenges of hypersonic threats.

Keywords:

Synthetic Image Generation, EOIR, Hypersonic Targets, ANSYS STK, ATR.

INTRODUCTION

1.1. BACKGROUND ON THE X-43A PROGRAM

NASA's X-43A program was part of the broader Hyper-X initiative, which marked a substantial breakthrough in hypersonic propulsion technology from multiple perspectives. This experimental vehicle successfully showed that a scramjet engine (short for supersonic combustion ramjet) could be combined with an airframe to achieve stable hypersonic flight. The program was launched in 1995-1996 to validate scramjet operability and generate data for designing future airbreathing hypersonic cruise platforms. The X-43A's successful flights in 2004 at Mach 7 and Mach 10 marked a significant advancement in hypersonic propulsion for US domestic capabilities [1], [2], [3].

The scramjet engine used gaseous hydrogen as fuel, with silane as its igniter. Its design was integrated seamlessly with the airframe, using the vehicle's forebody and afterbody as compression and expansion surfaces to maintain efficient airflow through the engine at hypersonic speeds.

Correspondence:

Slobodan Pajić

e-mail:

pajic428@gmail.com





Thermal protection was a crucial aspect of the design. Reinforced carbon-carbon was used for leading edges, and aluminium-enhanced thermal barrier tiles were used to shield the vehicle from extreme aerodynamic heating. The Mach 10 configuration incorporated enhanced thermal protections compared to the Mach 7 design, reflecting the higher thermal loads encountered at such speeds. The structure was composed of steel, titanium, and tungsten, materials chosen for their stiffness and thermal properties [3], [4]. The X-43A program's detailed aerodynamic and thermal performance datasets provided a foundational framework for modeling realistic EOIR target scenarios in this study.

The flight tests conducted as part of the Hyper-X program were very valuable. After a failed first flight in 2001 due to a booster malfunction, two successful flights in 2004 validated the scramjet engine's performance. On March 27, 2004, the X-43A achieved Mach 6.83; on November 16, 2004, it reached Mach 9.68, setting records for the fastest speeds ever achieved by an airbreathing vehicle. These tests confirmed that airframe-integrated scramjets could operate autonomously. These experiments produced data critical for validating computational models, aerodynamics, propulsion systems, and thermal management material technologies [2]- [4].

One of the significant challenges addressed during the program was achieving the necessary structural stiffness for the vehicle. Engineers relied heavily on finite element analysis to design a robust structure since intense bending moments were analyzed during flight. Vehicle weight was problematic since extensive usage of steel and aluminum was used to help with the structure rigidity. Despite these challenges, the X-43A transitioned smoothly between subsonic and supersonic combustion modes—a crucial feature of dual-mode scramjet engine design. This success was thanks to cutting-edge computational fluid dynamics (CFD) simulations and thorough wind tunnel testing [4].

The program also produced significant advancements in instrumentation and data acquisition. The X-43A was equipped with more than 500 sensors, including pressure taps and thermocouples, to gather precise data on the vehicle's aerodynamic pressure field, thermal characteristics, and structural performance. This array of instruments was used to deliver real-time data with telemetry equipment during the flights, allowing engineers to thoroughly analyze the vehicle's behavior and performance while flying at hypersonic speeds [3], [4].

The influence of the X-43A program goes far beyond its initial achievements. Proving that scramjet-powered hypersonic flight is possible paved the way for future innovations, such as reusable launch vehicles (RLVs) and hypersonic cruise missiles. Moreover, the data and insights gained have informed ongoing research into combined-cycle propulsion systems, such as Turbine-Based Combined Cycle (TBCC) and Rocket-Based Combined Cycle (RBCC) architectures [2]- [4].

1.2. CHALLENGES IN HYPERSONIC SCENARIO MODELLING

Hypersonic vehicles are complicated to detect and track because of their high speeds, unpredictable trajectories, and the intense heat generated by aerodynamic forces. Simulating these conditions using physics-based methods is particularly challenging for several reasons. Using synthetic data is a helpful way to recreate the extreme speeds and intense heat that hypersonic vehicles experience. It also reduces the need for expensive and potentially risky real-world data collection. The availability of real-world datasets for hypersonic scenarios is limited, mainly because of operational difficulties and strict confidentiality rules. Synthetic image generation allows the creation of datasets designed explicitly for automatic target recognition (ATR). Electro-optic infrared (EOIR) sensor models are key to making this process work effectively. Recreating critical elements of the X-43A flight test in synthetic environments establishes a robust foundation for further research or automatic target recognition (ATR) in the multispectral domain.

1.3. RESEARCH APPROACH AND OBJECTIVES

This study employs ANSYS STK tools to generate EOIR-emulated synthetic datasets, replicating the X-43A flight test scenario and expanding it with EOIR-rendered images. The approach minimizes computational complexity by simplifying the problem to flying objects without terrain while preserving operational relevance. The findings will demonstrate the potential of synthetic data in advancing hypersonic research.



2. RELATED WORK AND LITERATURE REVIEW

Synthetic data generation has revolutionized fields dependent on large-scale datasets by providing automated annotations, eliminating the time-consuming and costly manual labeling process. Despite its advantages, challenges such as domain adaptation and achieving realism persist when applying synthetic data to real-world scenarios [5], [6]. Two dominant approaches have emerged: domain randomization and photorealistic rendering. Domain randomization introduces diverse variations, such as lighting and object poses, to train robust neural networks [7]. Realistic rendering emphasizes visually accurate scenes with diverse environmental conditions like seasons and lighting [8], [9].

In defense applications, synthetic data has been instrumental in object detection and segmentation, with tools like DIRSIG modeling complex environments, including atmospheric and thermal conditions [6], [10]. Combining synthetic and real-world datasets, hybrid methods address the domain gap and enhance performance [6], [11]. These methodologies are well-suited for hypersonic scenarios, where challenges like high-speed targets and dynamic thermal signatures demand advanced synthetic datasets. Supported by tools such as STK EOIR, this study fills the gap by adapting synthetic data techniques to hypersonic detection [6], [10].

3. METHODOLOGY

The ANSYS STK EOIR module was the basis of this research. This tool provided high-fidelity radiometric sensor modeling, atmospheric effects simulation, and synthetic scenario scene generation. Using its capabilities, we simulated realistic hypersonic flight conditions and generated synthetic imagery to support developing and evaluating detection and tracking algorithms.

3.1. EXPERIMENTAL SETUP

Researchers developed the setup for studying the X-43A's flight, making sure it could accurately capture data during the vehicle's unpowered trajectory. Thanks to this detailed preparation, they were able to gather valuable insights into the vehicle's aerodynamic, thermal, and structural performance. As a result, researchers could thoroughly analyze and validate how the X-43A behaved while flying at hypersonic speeds [3], [4]. The mission employed a B-52 aircraft as the launch platform, following a pre-defined trajectory to achieve the specific altitude, velocity, and trajectory angles necessary to deploy the X-43A and its booster. The booster rocket carrying the X-43A was launched at the designated release point. Following the booster burn phase, a controlled separation was executed. Subsequently, the sustainer scramjet motor was activated, as illustrated in Figure 1 (a). Upon completion of the sustainer burn phase, depicted in Figure 1 (b), data about the unpowered flight phase was recorded precisely.

3.2. SCENARIO SETUP

We adopted a step-by-step approach to simulating the operational environment of X-43A hypersonic vehicle. The hypersonic vehicle was modeled with precise geometrical details, including surface materials characterized by their optical and thermal properties. The vehicle's path mimicked real-world flight dynamics, including fast acceleration, high-speed cruising, and gradual deceleration.

The simulation included atmospheric conditions that changed to match different operating scenarios, such as variations in temperature, pressure, humidity, and solar radiation. MODTRAN-based atmospheric models incorporated environmental effects such as transmission losses, scatter, and thermal path radiance. Scenarios included clear skies and degraded visual environments like

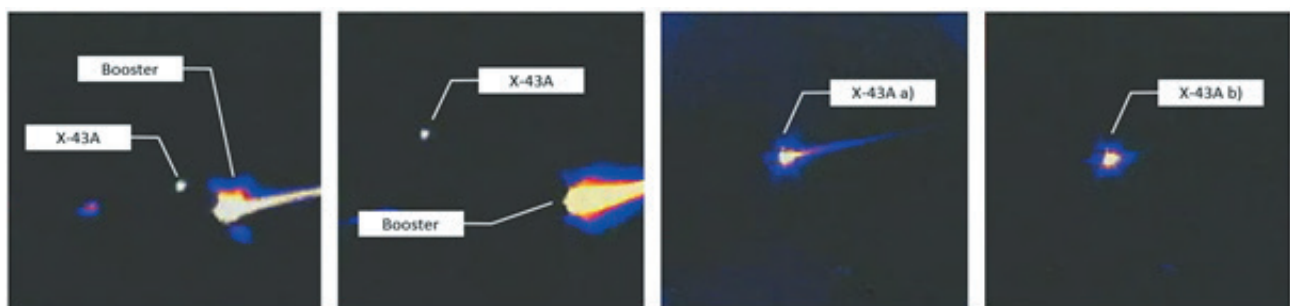


Figure 1. Separation and HALO-II experimental recording



cloud cover and haze. To reduce computational complexity, the scenario focused on aerial interactions without terrain modeling. Synthetic scenes were designed to include high-speed targets against varied spectral backgrounds, simulating EO/IR sensors of view at multiple altitudes and angles.

In Figure 2, we illustrate the simulation architecture developed using Ansys STK to model the flight and observation of the X-43A hypersonic aircraft during its unpowered flight phase.

The setup used a B-52 aircraft as the launch platform, with its flight path replicated from the actual experiment to accurately represent the mission's starting phase. The X-43A was designed with special attention to how separation occurred from the booster rocket, which marked the beginning of its sustained flight. The powered phase, including the sustain phase of the X-43A, was included from the scope of this analysis. The setup accounted for operational constraints by defining 20-35 km altitude ranges and velocity profiles from Mach 5 to Mach 10, reflecting real-world flight dynamics of hypersonic vehicles. The scenario depicted in Figure 3 included multiple observational platforms to monitor and analyze the unpowered flight: the HALO-II telemetry system with an EOIR camera platform and a satellite platform equipped with EOIR capabilities.

To replicate the experiment, the platforms were reconstructed, and a scenario was created in STK to simulate the unpowered flight phase in full detail.

3.3. SENSOR CONFIGURATIONS

Multiple EO/IR sensor configurations, HALO-II, Satellite, and Modified HALO-II, were tested, each parameterized with the spectral response, angular field of view, integration time, and resolution. Due to the unavailability of declassified data regarding the sensors, simulations were conducted to approximate and analyze the presented scenarios. A multi-sensor system was proposed, with sensors placed on simulated satellite and flight platforms. Each sensor was mounted on an independent gimbal, allowing it to collect different data points that gave a full picture of the target's radiometric signature. The Mod. HALO-II is a modified version of the previously analyzed HALO-II sensor, specifically designed and optimized to meet the requirements of ATR applications.

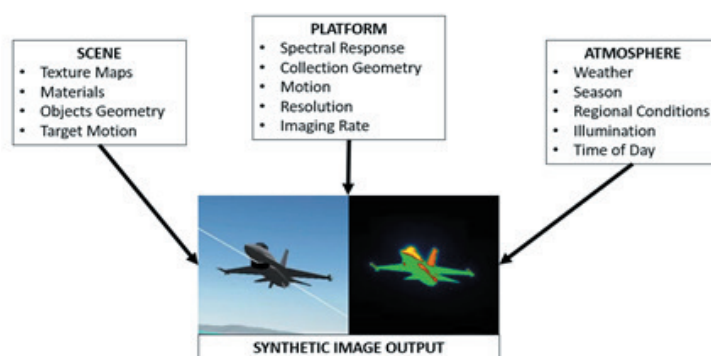


Figure 2. Ansys STK input components



Figure 3. Modelled scenario with 2D planar view of the mission as well as 3D view and FOV of the EOIR sensors



Table 1 shows that both HALO-II and Satellite function within a similar spectral range, which makes them well-suited for mid-wave infrared applications, including thermal imaging and environmental monitoring. The modified HALO-II, featuring a wider spectral range, improves its capacity to capture a variety of thermal and material signatures, thus making it an ideal choice for applications that require broader spectral coverage.

Table 2 presents the FOV analysis, revealing key differences in the scheme of the sensors. The HALO-II system features a geometric FOV of 0.4° , which allows it to cover a broader scene while still offering enough precision for fine angular details. On the other hand, the modified HALO-II has a similar number of pixels but delivers a slightly smaller effective FOV, which can be observed as striking a balance between resolution and coverage. Satellite, optimized for fine-resolution tasks, has a much smaller geometric FOV and instantaneous FOV, prioritizing high angular precision over coverage.

4. RESULTS

The Modulation Transfer Function (MTF) characteristics reveal distinct different sensor behaviors across the spatial frequency spectrum, as shown in Table 3. HALO-II exhibits a relatively steep decline in MTF at higher spatial frequencies, indicating a reduction in resolution and contrast for smaller targets. The decline in optics MTF is sharper compared to jitter and detector footprint MTFs, suggesting optical limitations as a key factor in high-frequency degradation. Mod. HALO-II shows a smoother MTF decline across frequencies, maintaining higher resolution at mid-to-low frequencies and being suitable for detecting moderately sized targets. Conversely, the Satellite demonstrates consistent MTF performance across spatial frequencies, reflecting robust optical and detector design tailored for uniform imaging tasks.

Table 1. Spectral Band Comparison

Sensor	Low Band Edge Wavelength (μm)	High Band Edge Wavelength (μm)
Satellite	3.0	5.5
HALO-II	3.0	5.5
Mod. HALO-II	2.5	6.0

Table 2. Geometric and Effective FOV Comparison

Attribute	Satellite	HALO-II	Mod. HALO-II
Horizontal Pixels	640.0	640.0	640.0
Vertical Pixels	640.0	640.0	640.0
Geometric FOV	7.8e-08	4.87e-05	4.87e-05
Horizontal Geometric FOV (deg)	0.016	0.4	0.4
Vertical Geometric FOV (deg)	0.016	0.4	0.4
Geometric Instantaneous FOV (Sterad)	1.9e-13	1.19e-10	1.19e-10
Horizontal Geometric Instantaneous FOV (mrad)	0.000436	0.0109	0.0109
Vertical Geometric Instantaneous FOV (mrad)	0.000436	0.0109	0.0109
Effective FOV (Sterad)	1.55e-05	7.12e-05	6.33e-05
Horizontal Effective FOV (deg)	0.226	0.484	0.456
Vertical Effective FOV (deg)	0.226	0.484	0.456
Effective Instantaneous FOV (Sterad)	3.79e-11	1.74e-10	1.55e-10
Horizontal Effective Instantaneous FOV (mrad)	0.00616	0.0132	0.0124
Vertical Effective Instantaneous FOV (mrad)	0.00616	0.0132	0.0124



Table 3. MTF Trends

Sensor	Spatial Frequency (cycle/mrad)	MTF Trends
HALO-II	Range (0 to 454 mrad)	Steep decline at high frequencies; optical limitations evident
Satellite	Range (0 to 235 mrad)	Consistent MTF performance; robust optical and detector design
Mod. HALO-II	Range (0 to 200 mrad)	Smoother decline, retaining resolution at mid-to-low frequencies

Table 4. Optical Specification Comparison

Specification	Satellite	HALO-II	Mod. HALO-II
Effective Focal Length (cm)	415.0	415.0	100.0
F Number	4.15	2.15	2.0
Diffraction Wavelength (μm)	4.25	4.25	2.5
Airy Disk Diameter (mrad)	0.0104	0.00537	0.0122
Rayleigh Resolution (mrad)	0.00519	0.00269	0.0061
Sparrow Resolution (mrad)	0.00425	0.0022	0.005
Ensquared Energy	0.005	0.685	0.77

Table 5. Sensor Responsivity and Noise Comparison

Metric	Satellite	HALO-II	Mod. HALO-II
Peak Irradiance Responsivity (cm²/W)	4.17e+19	2.12e+22	1.33e+21
Sensor NEI (W/cm²)	1e-15	1e-15	1e-15
Sensor SEI (W/cm²)	3e-12	3e-12	3e-12
Peak Radiance Responsivity (cm²·Sterad/W)	1.56e+9	3.49e+12	1.94e+11
Sensor NER (W·cm⁻²·Sterad⁻¹)	2.67e-05	6.06e-06	6.88e-06
Sensor SER (W·cm⁻²·Sterad⁻¹)	0.0802	0.0182	0.0206

HALO-II features an effective focal length of 415 cm and a relatively small F-number, resulting in a smaller Airy disk diameter and finer Rayleigh resolution, as shown in Table 4. This enables high spatial resolution and angular precision. Mod. HALO-II, with a shorter focal length and a slightly larger F-number, achieves a coarser resolution but offers broader field coverage and enhanced ensquared energy. Satellite focuses on mid-wave infrared imaging with a narrower F-number and comparable diffraction-limited resolution to HALO-II but significantly lower ensquared energy, prioritizing light collection efficiency for stable environmental conditions.

HALO-II excels in responsivity metrics, shown in Table 5, demonstrating high sensitivity to weak signals and suitability for low-light imaging. Mod. HALO-II, even while having slightly lower responsivity than HALO-II, compensates with superior noise performance and signal stability, making it suitable for capturing thermal contrasts. Satellite delivers reliable performance in both responsivity and noise management, striking an optimal balance between sensitivity and robustness for steady-state imaging tasks.

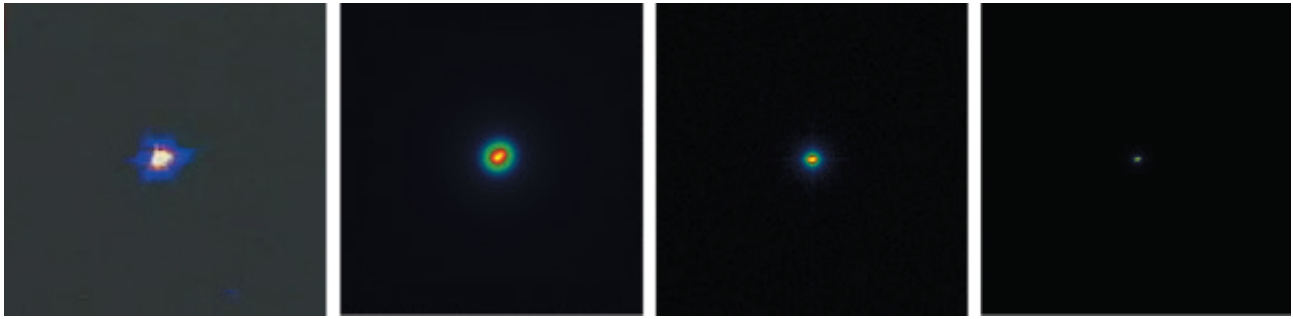


Figure 4. Comparison of experimental and synthetic data. The left panel shows the experimental data, followed by the Satellite data in the middle-left panel, HALO-II in the middle-right panel, and the Mod. HALO-II data on the far-right panel



Figure 5. ATR data set generated using Mod. HALO-II sensor model

This comparative analysis highlights the distinctive design philosophies of the three sensors depicted in Figure 4. Satellite demonstrates moderate performance across most metrics, optimized for mid-wave infrared applications requiring precise FOV control and high angular precision. HALO-II balances sensitivity, resolution, and coverage, making it suitable for versatile applications, particularly in low-light or broad-scene imaging. Mod. HALO-II excels in high-resolution imaging, superior angular precision, and broad spectral coverage, making it ideal for advanced ATR dataset generation and thermal imaging tasks.

The Mod. HALO-II sensor represents an optimized balance between resolution, field coverage, signal fidelity, and spectral versatility, making it a superior choice for ATR dataset generation, as depicted in Figure 5.

Building on the capabilities of the original HALO-II, Mod. HALO-II is designed to capture diverse, high-quality data, ensuring robust ATR model training and enabling algorithms to generalize effectively across

multiple operational scenarios. With its ability to handle diverse targets and environments, Mod. HALO-II supports the creation of comprehensive ATR datasets. These datasets capture key target attributes, including high-speed motion smear, thermal signatures, and environmental effects under varying conditions, such as clear skies, degraded visual environments (e.g., haze and cloud cover), and changes in the angular field of view and spectral response.



5. CONCLUSION

The use of synthetic datasets generated with ANSYS STK tools has proven to be an effective approach for modeling hypersonic vehicle detection scenarios, and ATR dataset generation, addressing significant data gaps in this field. These simulations have provided valuable insights into the operational behavior of EOIR sensors while also demonstrating an efficient workflow for ATR dataset creation. This methodology not only supports current research efforts but also establishes a reliable framework for future advancements in hypersonic detection and analysis. Future research should focus on the synergistic integration of synthetic datasets with experimental data to improve the adaptability and generalization capabilities of algorithms. By refining these methodologies, we can significantly enhance detection capabilities, paving the way for the seamless integration of synthetic data into AI-driven detection systems addressing the challenges posed by emerging hypersonic threats.

REFERENCES

- [1] C. R. McClinton, L. T. Nguyen, J. R. Sitz and L. V. Rausch, "Preliminary X-43 flight test results," *Acta Astronautica*, 2005.
- [2] P. T. Harsha, L. C. Keel, A. Castrogiovanni and R. T. Sherrill, "X-43A Vehicle Design and Manufacture," *AIAA/CIRA 13th International Space Planes and Hypersonics Systems and Technologies*, 2005.
- [3] C. McClinton, "X-43: Scramjet Power Breaks the Hypersonic Barrier," *NASA, Dryden Lecture*, 2006.
- [4] R. R. Kazmar, "Airbreathing Hypersonic Propulsion at Pratt & Whitney – Overview," *AIAA*, 2005.
- [5] S. Hinterstoisser, V. Lepetit, P. Wohlhart and K. Konolige, "On Pre-Trained Image Features and Synthetic Images for Deep Learning," *Computer Vision – ECCV 2018 Workshops*, p. pp 682–697, 2018.
- [6] S. Webster, M. Jeiran, K. Manser, N. Pinto and S. Harkrider, "Synthetic Imagery Generation," *NATO-OTAN, S&T Organization*, 2006.
- [7] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Bochooon and S. Birchfield, "Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018.
- [8] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent and R. Cipolla, "Understanding Real World Indoor Scenes With Synthetic Data," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [9] G. Ros, L. Sellart, J. Materzynska, D. Vazquez and A. M. Lopez, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [10] S. Han, A. Farfard, J. P. Kerekes, M. G. Gartley, E. J. Ientilucci, A. Savakis, C. Law, J. Parhan, M. Turek, K. Fieldhouse and T. Rovito, "Efficient generation of image chips for training deep learning networks," *Proceedings of SPIE - The International Society for Optical Engineering*, 2017.
- [11] J. Graybeal, J. Quartuccio and T. Du Bosq, "Immersive Simulation, Prototyping, and Evaluation of Infrared Sensor and Augmented Reality Technologies," *NATO-OTAN, S&T Organisation*, 2012.



UNSUPERVISED AND SEMI-SUPERVISED LEARNING TECHNIQUES IN CONTEMPORARY EDUCATIONAL APPLICATION

Veljko Aleksić*
[0000-0003-2337-1288]

University of Kragujevac,
Faculty of Technical Sciences in Čačak,
Čačak, Serbia

Abstract:

The paper examines the educational application of unsupervised and semi-supervised learning techniques. The comprehensive analysis evaluated diverse approaches, methods, and algorithms. Findings indicate that k-means clustering effectively differentiated student performance groups, while dimensionality reduction techniques offered valuable visualization capabilities for complex educational data. The semi-supervised learning paradigm demonstrated particular utility in environments characterized by abundant unlabelled data. The effectiveness of the presented analytical approaches significantly depends on data quality, appropriate algorithm selection, and domain expertise. As educational datasets grow increasingly complex, various computational methods will become essential in developing personalized learning, adaptive educational interventions, and innovative evidence-based teaching practices. The research contributes to the ever-evolving field of educational technology by systematically evaluating the strengths and limitations of various machine learning and artificial intelligence approaches, providing a foundation for future research.

Keywords:

Unsupervised Learning, Semi-Supervised Learning, Educational Technology, Algorithms.

INTRODUCTION

Artificial intelligence and machine learning are closely related terms often used interchangeably or as synonyms. However, they represent distinct concepts within the field of computer science. Artificial intelligence refers to the development of computational systems capable of performing tasks that typically require human intelligence, such as audio-visual perception, speech recognition, language translation, and even decision-making. The ultimate objective of artificial intelligence technology is the creation of virtual machines capable of thinking and acting in ways that simulate or surpass human cognitive abilities [1]. The domain of artificial intelligence encompasses a broad spectrum of approaches and techniques, including rule-based systems, expert systems, neural networks, and evolutionary algorithms [2]. These approaches are aimed at enabling computers to reason, learn, adapt, and solve problems in ways that emulate human intelligence [3]. It is noteworthy that artificial intelligence has, since its inception in the mid-20th century, progressed through cycles of increased and diminished enthusiasm and research investment [4].

Correspondence:

Veljko Aleksić

e-mail:

veljko.aleksic@ftn.kg.ac.rs





Despite these fluctuations, significant advancements have been achieved in recent decades, driven by developments in computer hardware and software, networking capabilities, the creation and accessibility of very large datasets (e.g., Big Data), and innovative algorithms. Machine learning should be perceived as a subfield of artificial intelligence focused on developing algorithms and statistical models that enable computers to learn and improve performance on specific tasks without explicit programming [5]. The basic idea is that computer software can learn from data, identify patterns, and make decisions or predictions with minimal human intervention [6]. Machine learning algorithms are generally categorized into three primary types: Supervised Learning (SL), Unsupervised Learning (UL), and Reinforcement Learning (RL). The field of machine learning has advanced significantly in recent years, particularly with the emergence of the Deep Learning (DL) model. DL should be perceived as a subdivision of machine learning that employs Artificial Neural Networks (ANNs) with multiple layers for learning and processing hierarchical data representations [7]. DL has achieved remarkable success in tasks such as image classification, speech recognition, and natural language processing. UL is a branch of machine learning directed toward uncovering hidden patterns, structures, and relationships within unlabelled data [8]. Unlike supervised learning, where algorithms learn from labeled input-output pairs, UL algorithms learn from the data with no explicit labels or target outputs to identify inherent structures, similarities, or clusters. UL algorithms possess significant potential in educational data mining and learning analytics, enabling the discovery of educational data hidden patterns, structures, and relationships. These algorithms can provide valuable insights into student behaviors, learning processes, and knowledge acquisition, thereby facilitating the design of personalized learning experiences, adaptive interventions, and curriculum enhancements [9] [10].

2. UNSUPERVISED LEARNING

Educational datasets often contain substantial quantities of unrelated data, including selection/click data, textual information, or multimedia content that is challenging to analyze and interpret without the assistance of UL algorithms. In the educational context, these algorithms are employed to address clustering challenges; grouping students based on their educational behaviors, preferences, or academic achievement levels; creating

recommendations for elective subjects and course selection; and organizing diverse educational resources [11]. UL algorithms can generally be classified into three categories:

- *Grouping (clustering) algorithms* partition input instances into unrelated discrete or overlapping groups, wherein instances within each group have greater similarity to one another compared to the instances in other groups. Prominent clustering algorithms are *k-means*, *Hierarchical Clustering*, and *Gaussian Mixture Models* (GMM) [12]. The k-means algorithm was applied to group students based on their academic performance and engagement levels in a blended learning environment [13], identifying four distinct student groups: high-performing, average, low-performing, and disengaged students, thereby assisting teachers in recognizing students at greater risk of academic failure and providing timely support and feedback;
- *Dimensionality reduction algorithms* transform high-dimensional input space into lower-dimensional representations while preserving the essential structure and information within the data [14]. This technique facilitates the visualization of complex datasets such as student interaction networks or conceptual maps, which can reduce computational complexity and enhance learning task performance. By reducing data dimensionality, educators can identify the most significant characteristics or variables that contribute to student learning outcomes and uncover hidden patterns or relationships among them. For instance, dimensionality reduction algorithms have demonstrated considerable usefulness in visualizing student learning trajectories within Massive Open Online Courses (MOOCs). Students who followed structured and linear pathways through MOOCs exhibited superior completion rates and higher overall results compared to those pursuing exploratory and non-linear trajectories [15]. Prominent dimensionality reduction algorithms are *Principal Component Analysis* (PCA), *t-Distributed Stochastic Neighbor Embedding* (t-SNE), and autoencoders [16].
- *Representation learning algorithms* derive new representations of input instances that detect fundamental variation factors or explanatory factors within the data. These learned representations can serve as input characteristics for supervised learning (classification or regression, both)



or unsupervised learning tasks (clustering or dimensionality reduction). In educational contexts, representation learning algorithms are used to extract meaningful and interpretable characteristics from educational data (e.g., student essays, forum posts, or lecture videos). The features identified and learned through these methods enhance automated essay scoring, knowledge tracking, and personalized recommendations or student feedback generation. Prominent representation learning algorithms are *Deep Belief Networks* (DBNs), *Restricted Boltzmann Machines* (RBMs), and variational autoencoders.

The previously mentioned algorithms enable researchers and educational practitioners to distill complex, multifaceted educational data into interpretable or structured representations that reveal fundamental learning patterns while maintaining statistical fidelity to the original information structure. Such approaches are essential as educational datasets grow both in size and complexity, necessitating sophisticated analytical techniques that can extract meaningful insights without overwhelming computational resources or analytical capacity. By identifying latent features and structures within educational data, UL algorithms provide educators with deeper insights into student learning patterns, conceptual understanding, and engagement behaviors that may not be immediately apparent through traditional analysis methods, thus significantly enhancing the potential for adaptive learning systems capable for dynamic respond to individual student needs and learning trajectories.

2.1. k-MEANS CLUSTERING METHOD

Clustering is one of the most frequently utilized UL techniques in educational data mining and learning analytics [17]. Clustering algorithms partition datasets into groups or clusters, ensuring that the data within a specific cluster exhibit greater similarity to each other than the data in other clusters. Similarity between the data is typically measured using metric distances, such as Euclidean distance or cosine similarity [18]. K-means clustering is a prominent partitional clustering algorithm that iteratively assigns data points to k clusters based on their similarity to cluster centroids. The algorithm starts with random initialization of k cluster centroids, followed by the two steps alternating execution: (1) assigning each data point to its nearest cluster center, and (2) updating centroids based on the average value

of data points assigned to each cluster. This process repeats until convergence is achieved, i.e., when assignments can no longer be changed [19]. In educational environments, the k-means clustering method is applied for grouping students based on their performance, learning behaviors, or engagement patterns [20]. For instance, this algorithm was employed to identify distinct student groups in MOOCs based on their interaction patterns with learning materials, quizzes, and discussion forums [21], revealing four clusters: active learners, passive learners, assessment-oriented learners, and dropout students. Clustering algorithms can also be utilized to group educational resources and evaluate their similarity in personalized education recommendations and curriculum design [22]. K-means algorithms are used for grouping educational videos based on content characteristics and usage patterns, thus enhancing personalized recommendations. These insights can inform teachers and other educators to adapt instructional design and support strategies for various student groups.

Hierarchical clustering can be performed by merging smaller clusters into larger ones (i.e., agglomerative clustering) or dividing larger clusters into smaller ones (i.e., divisive clustering). Agglomerative hierarchical clustering begins with each data point as a separate cluster and iteratively merges the closest clusters until a singular cluster is formed. Conversely, the divisive hierarchical clustering starts with all data points in one cluster and recursively divides clusters until each data point becomes a cluster [23]. In the educational context, hierarchical clustering is used to explore educational resources, curricula, and knowledge domain structures [24].

2.2. PRINCIPAL COMPONENT ANALYSIS

Dimensionality reduction represents another critical UL domain that aims to transform high-dimensional data into lower-dimensional representations while preserving the original essential structure and information. These techniques are particularly valuable for complex educational data visualization and exploration, such as student performance across multiple sequential assessments or the relationships between varying skill levels and knowledge components [25]. Principal Component Analysis (PCA) represents a frequently utilized linear dimensionality reduction technique that projects high-dimensional data onto a lower-dimensional subspace based on principal components that represent orthogonal directions capturing maximum variance within the



data [26]. The first principal component corresponds to the direction of the highest variance, with subsequent components progressively capturing smaller variance portions while maintaining orthogonality to preceding components. The retention of only the highest-ranked principal components can effectively reduce data dimensionality while preserving critical information. This technique is often used in educational data mining for student achievement data analysis and visualization, identification of the key factors influencing learning outcomes, and student behavioral patterns detection [27]. PCA can also be used for exploring relationships between student motivation, cognitive engagement, and academic performance in e-learning environments [28]. Numerous educational datasets contain a large number of characteristics and redundant or irrelevant information, impeding analytical performance and interpretability. PCA can serve as an effective mechanism for identifying the most informative characteristics and data dimensionality reduction, thus enhancing the efficiency and effectiveness of data mining and machine learning algorithms.

The widespread application of PCA in education is the interpretation of principal components because they are linear combinations of the original characteristics and may not have clear pedagogical or other specific meanings. In addition, PCA is sensitive to scaling or normalization of the original characteristics, as various pre-processing methods can lead to different results and interpretations. Educational datasets often contain a mixture of continuous, categorical, and binary characteristics, along with missing values, necessitating meticulous cleaning and transformation of the data before the analysis. As PCA assumes linear relationships between original characteristics it may fail to identify nonlinear structures or interactions in educational data. Nevertheless, PCA techniques continue to advance and are expected to play a significant role in data-driven decision-making, personalized learning, and evidence-based educational interventions.

2.3. *t*-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

t-Distributed Stochastic Neighbor Embedding (*t*-SNE) represents a nonlinear dimensionality reduction technique that is particularly efficient for visualizing 2D or 3D high-dimensional data. This methodology preserves local data structure by modeling similarities between data points in high-dimensional space and embedding them in low-dimensional space, thereby

minimizing the divergence between probability distributions representing similarity pairs in both high- and low-dimensional spaces [29]. In comparison to the PCA, this technique is more suitable for recognizing nonlinear relationships and preserving local data structure. However, it requires more processing time and is not the most efficient solution if the global data structure should be preserved [30]. By focusing on preserving local similarities between samples, this technique can reveal complex patterns and clusters in the data that may be “hidden” in linear projections. *t*-SNE requires very complex computing, exhibits sensitivity to hyperparameters selection, and encounters difficulties in preserving global data structure. In educational contexts, *t*-SNE is often used for visualizing and exploring complex student data, such as learning paths, skill mastery patterns, and, notably, misconception analysis. *t*-SNE is also utilized for visualizing and comparing student subgroups in educational data, thereby detecting various clustered groups of successful and unsuccessful students based on the identified key factors. When *t*-SNE is applied, the appropriate hyperparameters should be selected (e.g., controlling the balance between preserving local and global data structure and learning rate). As hyperparameters can significantly influence the quality and interpretability of the resulting visualization, different values are appropriate for various datasets and tasks [31]. Working with large educational datasets creates the problem of scalability, as the computational complexity of this technique quadratically increases with the number of instances. In addition, proper interpretation and evaluation require domain expertise, as although *t*-SNE can reveal interesting data patterns and clusters, the actual meaning or relevance of these patterns may not always be obvious in the educational context.

2.4. ASSOCIATION RULE MINING

Association Rule Mining (ARM) is a popular UL technique often used for discovering interesting relationships and patterns within large transactional or categorical datasets. This method aims to identify item sets that frequently co-occur and to generate associative rules describing appearance patterns and their interdependencies. ARM is implemented in two primary steps: (1) generating frequent item sets (e.g., the identification of sets that satisfy established constraints), and (2) generating rules from frequent item sets that meet specified criteria. The Apriori algorithm is usually combined with ARM as it uses a search strategy for generating frequent item sets. The algorithm is initiated by generating all



frequent item sets containing only one item by dataset scanning and counting. It then iteratively generates k -candidates from the frequent $(k-1)$ item sets, eliminates certain candidates based on the Apriori principle that every subset of a frequent item set must also be frequent, and recounts the remaining candidates by scanning the dataset. The process continues until all frequent item sets are generated. Finally, the algorithm generates “strong rules” from frequent item sets by calculating their confidence and verifying constraints [32]. ARM can also be performed via an FP-growth algorithm, that utilizes in-depth search and compressed tree structure for generating frequent item sets without generating the candidates. The algorithm first scans the dataset to identify all frequent one-item sets and sorts them in descending order of support. Then it rescans the dataset and constructs a tree by inserting each transaction as a tree path, while transaction items are arranged according to the frequency of one-item order. This is a compact dataset representation that enables efficient mining of frequent item sets. FP-growth algorithm recursively explores the tree by generating conditional pattern databases and constructs conditional trees for each frequent item. Frequent item sets are generated by combining frequent items in conditional trees with their corresponding prefix paths [33]. In educational contexts, ARM is used to discover patterns in student behavior, academic performance and success, and course selection structure. Having in mind that educational datasets often contain a large number of categorical variables (e.g., demographic characteristics, activity logs, learning activities, grades, etc.), they are suitable for this technique, allowing teachers and researchers to gain insights into the relationships between different variables and generate recommendations for teaching and learning improvement [34].

3. SEMI-SUPERVISED LEARNING

Semi-supervised learning (SSL) represents a machine learning paradigm that utilizes both labeled and unlabelled data for predictive model performance enhancement. Contrary to the unlabelled, labeled data are usually scarce or expensive to obtain. SSL aims to exploit the hidden structure and information within unlabelled datasets to improve the learning process and reduce the reliance on labeled sets [35]. The SSL model assumes that the distribution of unlabelled data carries valuable information about the basic data structure and can assist in constraining and directing predictive model learning. The smoothness assumption states that data points

situated in proximity to one another in the input space are likely to have similar output values or belong to the same class. Leveraging this assumption, SSL algorithms can propagate information from labeled examples to unlabelled instances and enhance the generalization performance of learned models [36]. SSL has several major approaches:

- *Self-training* constitutes an iterative process wherein a classifier is initially trained on labeled data and subsequently used to predict labels for unlabelled instances. The most confident predictions are incorporated into the labeled dataset, and the classifier is retrained on the expanded dataset. The process repeats until a stopping criterion is achieved or all unlabelled instances are labeled [37]. Self-training is applied in text content classification, image recognition, and speech recognition;
- *Co-training* assumes that the input characteristics can be partitioned into two or more independent and sufficient views, each capable of predicting the target variable. The algorithm trains separate classifiers of each view using labeled data and then uses the most confident predictions from each classifier to expand the labeled set for other views. The classifiers are iteratively retrained on expanded datasets until a stopping criterion is met [38]. This approach is used for web page classification, natural language processing, and bioinformatics;
- *Tri-training* extends co-training without requiring multiple views. Instead, three classifiers are trained on different labeled data samples obtained via bootstrap techniques, and the predictions of two classifiers are used to label the unlabelled instances for the third classifier. The process repeats for each classifier until a stopping criterion is achieved [39]. Tri-training is efficient for text categorization, image classification, and software defect prediction;
- *Graph-based method* represents data as a graph where nodes correspond to labeled and unlabelled instances, and edges reflect similarity between instances. The main idea is to propagate the label information from the labeled nodes to unlabelled ones based on the graph structure, assuming that the connected nodes have similar labels. This method is used in computer vision, natural language processing, and recommendation systems [40];



- *Generative models* analyze the joint distribution of input characteristics and target variables using both labeled and unlabelled data. The primary concept involves training a generative model to produce realistic samples from the data distribution and to employ it to infer the missing labels of unlabelled instances. Generative models are based on various frameworks, including Gaussian Mixture Models, Hidden Markov Models, and Variational Autoencoders [41], and are applied in text classification, speech recognition, and image generation.

SSL has numerous educational applications. It can be used for predicting student grades in distance learning courses utilizing both labeled and unlabelled data to enhance prediction accuracy [42], or creating personalized recommendations within e-courses employing available grades and implicit feedback from student interactions with the learning management system [43]. SSL performance primarily depends on the quality and representativeness of unlabelled data, as well as the validity of assumptions regarding data distribution. If unlabelled data sets are biased or contain instances from different domains or distributions, SSL will yield poor performance [44].

4. CONCLUSION

The paper presented the multifaceted landscape of unsupervised and semi-supervised learning techniques in educational applications. The investigation of clustering algorithms, dimensionality reduction methods, representation learning approaches, and association rule mining has revealed their significant potential for extracting meaningful patterns from complex educational datasets. *K-means* clustering has demonstrated substantial utility in identifying distinct student groups based on performance and engagement metrics, enabling educators to implement targeted interventions. PCA and t-SNE have proven valuable for visualizing high-dimensional educational data, though each presents distinct advantages regarding linear versus nonlinear relationship detection. ARM techniques have facilitated the discovery of meaningful correlations between student behaviors and academic outcomes, providing actionable insights for instructional design. SSL approaches are particularly promising in educational contexts in case labeled data are scarce and unlabelled data are abundant. Self-training, co-training, tri-training, graph-based methods, and generative models each of-

fer unique capabilities for propagating label information across datasets, potentially enhancing the accuracy of predictive models while reducing reliance on extensively labeled examples. Nevertheless, the efficacy of these techniques critically depends on data quality, representativeness, and the validity of underlying assumptions regarding data distribution. Careful algorithm selection, hyperparameter tuning, and domain expertise remain essential for ensuring optimal results in educational applications.

As educational datasets continue to grow in both size and complexity, the presented sophisticated analytical approaches will play increasingly vital roles in personalized learning, adaptive learning, and evidence-based educational design. Future research will focus on enhancing the interpretability of numbered techniques, particularly for educational stakeholders without extensive data science expertise, thereby maximizing their practical usefulness in improving the educational process.

5. ACKNOWLEDGEMENTS

This study was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, and these results are parts of the Grant No. 451-03-136/2025-03/200132 with University of Kragujevac, Faculty of Technical Sciences Čačak.

REFERENCES

- [1] H. Pedersen, "Existentialism and Artificial Intelligence in the 21st Century," in *The Routledge Handbook of Contemporary Existentialism*, Routledge, 2024, pp. 36-45.
- [2] N. Vlassis, *A concise introduction to multiagent systems and distributed artificial intelligence*, Morgan & Claypool Publishers, 2007.
- [3] R. Verganti, L. Vendraminelli and M. Iansiti, "Innovation and design in the age of artificial intelligence," *Journal of product innovation management*, vol. 37, no. 3, pp. 212-227, 2020.
- [4] L. Floridi, "AI and its new winter: From myths to realities," *Philosophy & Technology*, vol. 33, pp. 1-3, 2020.
- [5] Z. H. Zhou, *Machine learning*, Springer Nature, 2021.
- [6] S. Vieira, W. H. L. Pinaya and A. Mechelli, "Introduction to machine learning," in *Machine learning*, Academic Press, 2020, pp. 1-20.
- [7] T. J. Sejnowski, *The deep learning evolution*, MIT Press, 2018.



- [8] D. Valkenborg, A. J. Rousseau, M. Geubbelmans and T. Burzykowski, "Unsupervised learning," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 163, no. 6, pp. 877-882, 2023.
- [9] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 10, no. 3, p. e1355, 2020.
- [10] A. Dutt, M. A. Ismail and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991-16005, 2017.
- [11] E. Cam and M. E. Ozdag, "Discovery of Course Success Using Unsupervised Machine Learning Algorithms," *Malaysian Online Journal of Educational Technology*, vol. 9, no. 1, pp. 26-47, 2021.
- [12] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651-666, 2010.
- [13] A. Dutt, S. Aghabozrgi, M. A. B. Ismail and H. Mahrooian, "Clustering algorithms applied in educational data mining," *International Journal of Information and Electronics Engineering*, vol. 5, no. 2, p. 112, 2015.
- [14] A. K. Rastogi, S. Taterh and B. S. Kumar, "Dimensionality Reduction Algorithms in Machine Learning: A Theoretical and Experimental Comparison," *Engineering Proceedings*, vol. 59, no. 1, p. 82, 2023.
- [15] Z. Alj, A. Bouayad and M. O. Cherkaoui Malki, "Dimensionality reduction for predicting students dropout in MOOC," in *International Conference on Digital Technologies and Applications*, Cham: Springer Nature Switzerland, 2023.
- [16] Y. Bengio, A. Courville and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798-1828, 2013.
- [17] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. D. F. Costa and F. A. Rodrigues, "Clustering algorithms: A comparative approach," *PloS one*, vol. 14, no. 1, p. e0210236, 2019.
- [18] G. Gan, C. Ma and J. Wu, *Data clustering: theory, algorithms, and applications*, Society for Industrial and Applied Mathematics, 2020.
- [19] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716-80727, 2020.
- [20] A. Moubayed, M. Injadat, A. Shami and H. Lutfiyya, "Student engagement level in an e-learning environment: Clustering using k-means," *American Journal of Distance Education*, vol. 34, no. 2, pp. 137-156, 2020.
- [21] M. Khalil and M. Ebner, "Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories," *Journal of computing in higher education*, vol. 29, pp. 114-132, 2017.
- [22] V. Balakrishnan and E. Lloyd-Yemoh, "Stemming and lemmatization: A comparison of retrieval performances," *Lecture notes on software engineering*, vol. 2, no. 3, p. 262, 2014.
- [23] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn and C. Mathieu, "Hierarchical clustering: Objective functions and algorithms," *Journal of the ACM (JACM)*, vol. 66, no. 4, pp. 1-42, 2019.
- [24] A. Triayudi and I. Fitri, "A new agglomerative hierarchical clustering to model student activity in online learning," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 17, no. 3, pp. 1226-1235, 2019.
- [25] B. K. Tripathy, S. Anveshritaa and S. Ghela, "Comparative Analysis of Dimensionality Reduction Techniques," in *Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization*, Routledge, 2021, p. 137-149.
- [26] M. Greenacre, P. J. Groenen, . T. Hastie, A. I. d'Enza, A. Markos and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, 2022.
- [27] G. Feng, M. Fan and Y. Chen, "Analysis and prediction of students' academic performance based on educational data mining," *IEEE Access*, vol. 10, pp. 19558-19571, 2022.
- [28] H. Hershcovits, D. Vilenchik and K. Gal, "Modeling engagement in self-directed learning systems using principal component analysis," *IEEE Transactions on Learning Technologies*, vol. 13, no. 1, pp. 164-171, 2019.
- [29] M. Balamurali, "t-Distributed Stochastic Neighbor Embedding," in *Encyclopedia of Mathematical Geosciences*, Cham: Springer International Publishing, 2021, pp. 1-9.
- [30] M. Wattenberg, F. Viégas and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, p. e2, 2016.
- [31] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *The journal of machine learning research*, vol. 15, no. 1, pp. 3221-3245, 2014.
- [32] D. M. D. Angeline, "Association rule generation for student performance analysis using apriori algorithm," *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, vol. 1, no. 1, pp. 12-16, 2013.
- [33] L. Zhiyuan, "Research on Association Rule Mining Algorithm Based on FP-tree," in *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, IEEE, 2024.
- [34] L. Zhou and S. Yau, "Association rule and quantitative association rule mining among infrequent items," in *Proceedings of the 8th international workshop on Multimedia data mining:(associated with the ACM SIGKDD 2007)*, ACM, 2007.



- [35] Z. Chen and B. Liu, "Lifelong Semi-supervised Learning for Information Extraction," in *Lifelong Machine Learning*, Cham: Springer International Publishing, 2017, pp. 81-88.
- [36] Z. Miao, Y. Zhang, X. Piao, Y. Chu and B. Yin, "Region feature smoothness assumption for weakly semi-supervised crowd counting," *Computer Animation and Virtual Worlds*, vol. 34, no. 3-4, p. e2173, 2023.
- [37] J. Jeong, S. Lee and N. Kwak, "Self-Training using Selection Network for Semi-supervised Learning," in *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, 2020.
- [38] C. Yabi and Z. Yongzhao, "Co-training semi-supervised active learning algorithm based on noise filter," in *2009 WRI Global Congress on Intelligent Systems* (Vol. 3), IEEE, 2009.
- [39] Y. Zhang and S. Yan, "Semi-supervised active learning image classification method based on tri-training algorithm," in *2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIS)*, IEEE, 2020.
- [40] K. Avrachenkov and M. Dreveton, "Graph-based semi-supervised learning," in *Statistical Analysis of Networks*, Now Publishers, 2022, p. 5.
- [41] H. Yoon and H. Kim, "Label-noise robust deep generative model for semi-supervised learning," *Technometrics*, vol. 65, no. 1, pp. 83-95, 2023.
- [42] G. Kostopoulos, S. Kotsiantis, N. Fazakis, G. Koutsoukos and C. Pierrakeas, "A semi-supervised regression algorithm for grade prediction of students in distance learning courses," *International Journal on Artificial Intelligence Tools*, vol. 28, no. 04, p. 1940001, 2019.
- [43] A. Elbadrawy and G. Karypis, "Domain-aware grade prediction and top-n course recommendation," in *Proceedings of the 10th ACM conference on recommender systems*, ACM, 2016.
- [44] G. Abdelrahman, Q. Wang and B. Nunes, "Knowledge tracing: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1-37, 2023.



BLOCKCHAIN-BASED PLATFORM FOR TRACKING STUDENTS' ACHIEVEMENTS

Amar Mujezinović,
[0009-0001-9388-2350]

Aleksandar Joksimović*,
[0009-0008-5711-7636]

Dušan Kostić,
[0009-0000-3432-6074]

Miloš Jolović,
[0009-0003-6580-2039]

Milica Simić
[0000-0002-6870-2303]

University of Belgrade,
Faculty of Organizational Sciences,
Belgrade, Serbia

Abstract:

The increasing need for secure, transparent, and verifiable academic credentials has highlighted the limitations of conventional digital certificates, which are often static, hard to validate, and prone to fraud. To address this gap, this paper explores blockchain technology and Non-Fungible Tokens (NFTs) as innovative solutions for academic credentialing. Unlike traditional credentials, NFTs are unique, immutable digital assets stored on a blockchain, enabling decentralized verification and secure ownership.

This study proposes a blockchain-based platform that enables universities to issue NFT (Non-Fungible Token) tokens to students upon completion of coursework, participation in extracurricular activities, and acquisition of specific skills. These tokens are stored on a decentralized ledger and form a tamper-proof record of student achievements accessible to employers, institutions, and other stakeholders.

The paper introduces the "Achievables" business model, which promotes collaboration between universities, students, and employers. A prototype implementation using Ethereum's ERC-721 standard demonstrates the system's core features. Challenges such as scalability, legal frameworks, and institutional adoption are also addressed. The findings suggest that blockchain-based academic credentialing can enhance transparency, student motivation, and employability, offering a secure and learner-centered approach to achievement tracking.

Keywords:

Non-Fungible Tokens (NFTs), Higher Education, Digital Credentials, Academic Achievement Tracking, Decentralized Learning.

INTRODUCTION

Modern higher education is embracing digital transformation to enhance student competitiveness through personalized portfolios and verified records of academic and extracurricular achievements. By leveraging blockchain and NFTs, institutions can issue secure and transparent digital badges that represent milestones, such as course completion, research presentations, or competition awards. These tokens are stored in digital wallets and linked to a decentralized ledger, allowing easy verification by employers, scholarship providers, and other stakeholders. This approach replaces traditional paper certificates with tamper-proof digital credentials, while also providing universities with valuable insights into student engagement, curriculum effectiveness, and long-term outcomes.

Correspondence:

Aleksandar Joksimović

e-mail:

aleksandar.joksimovic@elab.rs





2. LITERATURE REVIEW

While blockchain's initial applications were concentrated in the financial sector, its potential in education is expanding rapidly, encompassing digital diplomas, e-learning, and decentralized accreditation systems [1]. Universities are increasingly recognizing the efficiency of blockchain-powered solutions, leading to the development of platforms such as EduCTX, which is designed to provide a globally recognized and transparent system for managing academic credits [2].

2.1. BLOCKCHAIN TECHNOLOGIES

Blockchain is a decentralized and immutable digital ledger that records transactions securely across multiple computers. It eliminates intermediaries, enhances security, and ensures transparent data verification. Each transaction is grouped into a block and linked cryptographically, forming a tamper-proof chain of records [3]. Consensus mechanisms, such as Proof of Work (PoW) and Proof of Stake (PoS), govern the validation of transactions. While PoW relies on computational puzzles, PoS selects validators based on token holdings, offering a more energy-efficient alternative. Blockchain has applications in finance, supply chain management, healthcare, and identity verification. It enables decentralized financial transactions, enhances supply chain transparency, and secures patient records. However, challenges such as scalability, regulatory uncertainty, and interoperability between networks remain. [4]. Ongoing advancements, including Layer 2 scaling solutions and hybrid blockchain models, aim to address these issues. With continued development, blockchain is expected to revolutionize digital transactions and data management across various industries. [5].

2.2. NFTS AND NFT MARKETPLACES IN THE CONTEXT OF EDUCATION

Non-Fungible Tokens (NFTs) are unique digital assets stored on a blockchain that verify ownership of digital or physical items. Unlike cryptocurrencies, which are interchangeable, each NFT carries specific metadata that makes it distinct and non-replicable [6]. This makes NFTs particularly suitable for representing academic credentials, as each educational achievement can be encoded with personalized metadata such as course name, grade, instructor, and timestamp.

NFT marketplaces, including OpenSea, Rarible, and Foundation, illustrate the technical and economic frameworks for decentralized asset management [7][8]. Although initially intended for creative industries, their structure inspires the development of educational platforms where students can securely store, manage, and share their credentials. By adopting the smart contract mechanisms used in mainstream NFT platforms to automate royalty payments, educational credentials can be issued and tracked securely and efficiently.

Despite their rapid adoption, NFTs face challenges such as price volatility, legal ambiguities, and concerns over environmental impact due to blockchain energy consumption [9]. However, in the context of education, NFTs offer promising advantages in creating transparent, tamper-proof, easily shareable, and verifiable personal records across institutions and platforms.

2.3. APPLICATION OF BLOCKCHAIN AND NFT IN HIGHER EDUCATION

The potential of blockchain and NFTs in education is expanding rapidly, introducing new ways to manage academic credentials, digital learning, and institutional governance. Traditional systems for storing academic records rely on centralized databases, which are vulnerable to fraud, data loss, and unauthorized modifications. Blockchain's decentralized nature ensures that academic credentials are securely stored and verifiable by multiple stakeholders, including universities, employers, and government agencies. One of the key applications of blockchain in education is the issuance of digital diplomas and certifications [10]. Platforms such as EduCTX integrate blockchain with the European Credit Transfer and Accumulation System (ECTS), allowing students to store academic credits in a tamper-proof system. This reduces reliance on paper-based documentation and eliminates intermediaries in credential verification [2]. Blockchain-powered Decentralized Autonomous Organizations (DAOs) offer students and faculty members an innovative approach to managing academic policies, learning resources, and institutional governance. Through smart contracts, these systems ensure transparency and inclusivity in decision-making, allowing students to participate in curriculum development and academic management [10].

Smart contracts also automate administrative processes, including degree verification, student enrollment, and the distribution of financial aid. Automated systems reduce operational costs, mitigate fraud risks,



and enhance institutional efficiency. Blockchain solutions such as EduRSS enable secure storage and exchange of student achievements, preventing data manipulation and unauthorized access. [11]. NFTs further revolutionize education by offering digital credentialing. Instead of traditional degrees, universities can issue NFT-based digital badges that represent specific skills or competencies acquired through coursework or training programs. These verifiable credentials provide an immutable record of student achievements that can be shared directly with employers and institutions. Beyond credentialing, NFTs facilitate the monetization of educational content. Professors and researchers can tokenize digital textbooks, research papers, and online courses to ensure intellectual property protection. Blockchain-based NFT marketplaces enable the direct distribution of educational materials, eliminating intermediaries and enhancing accessibility and revenue potential for educators. NFT-backed scholarships and funding models present new opportunities for student financial aid. Universities and donors can issue NFT-backed grants, ensuring transparency in fund allocation and supporting students through blockchain-verified sponsorships [12], [13]. Despite these advantages, challenges such as scalability, legal recognition, and environmental concerns remain. Blockchain networks require substantial infrastructure, and the lack of interoperability between platforms remains a significant challenge. Additionally, regulatory frameworks must be developed to standardize NFT-based academic credentials [14].

While prior research and platforms such as EduCTX [2] and EduRSS [11] demonstrate the feasibility of blockchain in academic credentialing, they typically rely on static, institution-driven certificates. These systems lack support for dynamic, skill-specific NFTs and offer limited personalization or student involvement in the credentialing process. This paper addresses that gap by proposing a learner-centered, NFT-based platform that combines academic and extracurricular achievements into a unified, verifiable portfolio. By doing so, it introduces a more modular, transparent, and practical approach to digital credentialing.

3. BUSINESS MODEL FOR A BLOCKCHAIN-BASED PLATFORM FOR TRACKING STUDENTS' ACHIEVEMENTS

The "Achievables" business model introduces a blockchain-powered platform that enables universities to issue unique NFT collections as verifiable digital credentials for students' achievements. Rather than focusing solely on GPA or degrees, the platform highlights evolving skill profiles, reflecting students' academic progress and extracurricular activities. Universities create custom NFT badges, which are linked to specific courses, workshops, and formal events. Institutions submit enrollment lists, top-performing students, and graduates, prompting the automatic minting and issuance of NFTs. These tokens serve as secure, decentralized proof of accomplishments. Additionally, students can request personalized NFTs to showcase independent achievements, such as research publications or specialized training. To better conceptualize the proposed platform and its operational structure, the following section utilizes Business Model Canvas. This approach outlines the core elements needed to build, deliver, and capture value in a blockchain-based academic credentialing ecosystem.

- **Business idea:** A platform (named "Achievables") that integrates blockchain technology into education by offering universities unique NFT collections as a form of student recognition for specific academic achievements.
- **Products and services:** Creation of NFT collections for universities, a platform that highlights university NFT collections with a searchable database for hiring companies.
- **Partners:** Universities; Companies looking to hire young talent; Students; Designers
- **Key activities:** Creating NFT collections for universities; Awarding students NFTs upon reaching specific achievements; Handling unique NFT requests for student profiles; Showcasing NFTs with student profiles on the platform; Collaborating with hiring companies via a searchable skills database; Updating the database annually with universities; Managing student complaints regarding database errors.
- **Value propositions:** Comprehensive portfolio building for students; Customized learning pathways; Enhanced student engagement; Empowerment of non-traditional learners; Competitive advantage for new graduates; Streamlined talent



acquisition for employers; Global recognition of academic achievements; Data-driven insights for universities; Promotion of blockchain literacy

- **Relationship with users:** Community building; Intuitive, user-centric design; Effective communication and feedback mechanisms; Personalization; Transparency; Consistent value delivery
- **Customer segments:** Universities and educational institutions; Students; Employers and hiring teams; Providers of educational services; Regulatory bodies and accreditation agencies; Investors and stakeholders
- **Key resources:** Financial capital; Human capital; Technological infrastructure; Blockchain technology; Data management systems; Regulatory compliance
- **Channels:** Online platform; Digital marketing; Partnerships with educational institutions
- **Cost structure:** Technology development and maintenance; Blockchain integration costs; Employee salaries; Marketing and promotion; Legal fees; Office and administration expenses
- **Revenue Streams:** Revenue from partnership and collaboration fees with educational institutions and companies

The "Achievables" business model, as described above, is structured around subscription fees from universities integrating the platform into their academic administration, as well as membership fees from companies accessing the student talent pool. Employers can use advanced search features to identify candidates with specific skills or match them to NFT-verified credentials. Students benefit from enriched academic experiences,

improved career prospects, and increased motivation to engage beyond the standard curriculum. The platform also collaborates with digital artists to design institution-branded NFT graphics. Additionally, small transaction or service fees apply when students request personalized NFTs for extracurricular achievements. This ecosystem fosters close collaboration between universities, employers, and students while actively promoting blockchain literacy in education.

4. DEVELOPED A BLOCKCHAIN-BASED PLATFORM FOR TRACKING STUDENTS' ACHIEVEMENTS

Figure 1 displays the prototype homepage of the platform, built with React, highlighting its core features and user interface. The platform is built on Ethereum's ERC-721 standard for NFTs and features a wallet-based login, a user dashboard showcasing earned NFT credentials, and administrative panels for token creation and management. It supports essential operations such as minting new tokens, verifying ownership, and recording transactions through a dedicated smart contract referred to as *NFTMarketplace.sol*. To ensure compatibility with major digital wallets and marketplaces, such as MetaMask, the platform integrates key contract modules, including *ERC721.sol*, *ERC721Enumerable.sol*, *IERC721.sol*, *IERC721Enumerable.sol*, and *IERC721Metadata.sol*. Instead of presenting the Solidity code directly, Listing 1 provides a pseudocode version of the minting logic, abstracting technical implementation details while preserving the functional structure. This approach demonstrates how the platform issues NFT-based academic credentials while promoting good security practices [15].

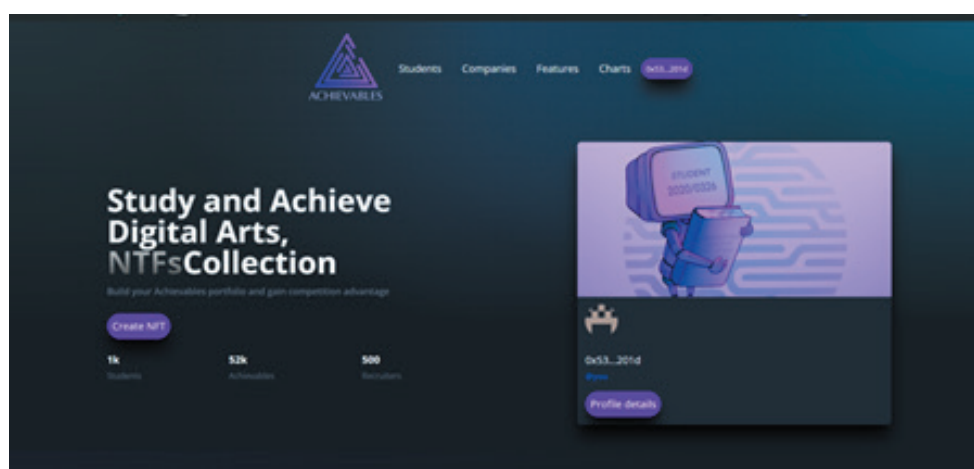


Figure 1. Platform for Tracking Students' Achievements



```

FUNCTION mintNFT(title, description, metadataURI, salesPrice, paymentAmount):

    IF metadataURI already exists:
        REJECT transaction ("This NFT is already minted")

    IF paymentAmount < mintingCost:
        REJECT transaction ("Ether too low in minting")

    IF sender IS contractOwner:
        REJECT transaction ("Sales not allowed")

    TRANSFER royalty TO artist
    TRANSFER remainingAmount TO platformOwner

    CREATE transactionRecord WITH metadata:
        - tokenId
        - sender
        - paymentAmount
        - salesPrice
        - title
        - description
        - metadataURI
        - timestamp

    EMIT sale event WITH transactionRecord
    MINT new NFT with metadataURI to sender
    RECORD sender AS NFT owner
    MARK metadataURI AS used

```

Listing 1. Mint NFTs function

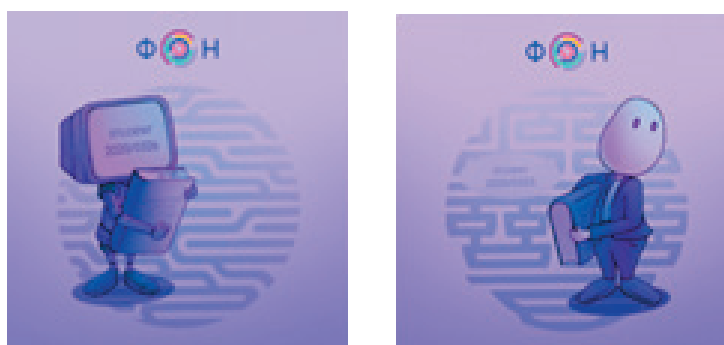


Figure 2. IT Student and Management Student

This function ensures that only valid, unique tokens are issued. It performs key validation checks, calculates and distributes royalties, creates an on-chain transaction record, and safely mints the token to the student's wallet. Each NFT includes descriptive metadata such as the student's name, the course or achievement represented, the issuing institution, and the time of issuance. By abstracting these operations into a smart contract, the system maintains transparency, traceability, and automation in the credentialing process. The system's decentralized design reduces administrative workload while enabling students to control and share their credentials independently.

When a student first enrolls in a university that is part of the system, they receive an enrollment NFT that attests to their official registration on the platform.

Figure 2 presents sample NFT tokens issued to students upon enrollment in their respective study programs. These enrollment NFTs serve as the initial credentials in a student's blockchain-based academic record. Each token includes metadata that identifies the study track (e.g., IT or Management), the enrollment year, and the issuing institution. This mechanism replaces traditional registration confirmations with verifiable, tamper-proof credentials anchored on the blockchain.



By integrating institutional branding and metadata, these NFTs ensure authenticity while laying the foundation for subsequent academic achievements within the platform.

As students progress through their studies, they may receive additional NFTs upon completion of specific courses or projects. Each token contains metadata such as course name, learning objectives, instructor identity, and grade, thereby forming a dynamic and detailed academic portfolio over time. Beyond academic milestones, the platform also supports the issuance of NFTs for extracurricular accomplishments. These may include participation in hackathons, boot camps, or independent research, with tokens encoding verified information about the acquired skills and the context in which they were gained. This functionality enables institutions to validate and reward a broader range of student development, encompassing both formal and informal learning outcomes.

5. CONCLUSION

By integrating blockchain and NFT technologies with conventional models of higher education, this platform conceptually illustrates a next-generation credentialing system that addresses several chronic pain points in academic and professional settings. The decentralized blockchain enables tamper-proof record-keeping, thus reducing the administrative burden of verifying records and minimizing credential fraud. NFTs enable granularity to represent each student's unique accomplishments, offering employers more insight into individual competencies. While eventual roadblocks include transaction costs on clogged networks, differing institutional policies, and the need for standardized data formatting, the current development of blockchain infrastructure and the broader adoption of digital assets suggest that such barriers can be gradually eroded. Additionally, such a system could enhance student motivation, as they earn tokens for diverse skill sets that are reflected in their final transcripts. Universities, in return, get a secure data platform that bolsters their reputation for innovation while providing them with real-time snapshots of student achievement. As higher education increasingly embraces digital transformation, blockchain-based achievement-tracking platforms are going to be at the heart of how credentials get stored, shared, and verified.

REFERENCES

- [1] F. Tschorsch and B. Scheuermann, "Bitcoin and beyond: A technical survey on decentralized digital currencies," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 3, pp. 2084–2123, Jul. 2016, doi: 10.1109/COMST.2016.2535718.
- [2] M. Turkanović, M. Hölbl, K. Košič, M. Heričko, and A. Kamišalić, "EduCTX: A blockchain-based higher education credit platform," *IEEE Access*, vol. 6, pp. 5112–5127, Jan. 2018, doi: 10.1109/ACCESS.2018.2789929.
- [3] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," Accessed: Mar. 10, 2025. [Online]. Available: www.bitcoin.org
- [4] Z. Bogdanović, B. Radenković, M. Despotović-Zrakić, D. Barać, A. Labus, and T. Naumović, "Blockchain technologies: current state and perspectives," *Zbornik radova Međunarodne naučne konferencije o digitalnoj ekonomiji DIEC*, vol. 2, no. 2, pp. 1–12, 2019.
- [5] A. Gangwal, H. R. Gangavalli, and A. Thirupathi, "A survey of Layer-two blockchain protocols," *Journal of Network and Computer Applications*, vol. 209, p. 103539, Jan. 2023, doi: 10.1016/J.JNCA.2022.103539.
- [6] M. Simić, K. Šikman, M. Vrljanac, V. Despotović, and M. Despotović-Zrakić, "Methodology of Creating NFT Fashion Projects," *Facta Universitatis, Series: Electronics and Energetics*, vol. 37, no. 1, pp. 29–52, 2024, doi: 10.2298/FUEE2401029S.
- [7] Q. Wang, R. Li, Q. Wang, and S. Chen, "Non-Fungible Token (NFT): Overview, Evaluation, Opportunities and Challenges," May 2021, Accessed: Mar. 10, 2025. [Online]. Available: <https://arxiv.org/abs/2105.07447v3>
- [8] K. Vasan, M. Janosov, and A. L. Barabási, "Quantifying NFT-driven networks in crypto art," *Scientific Reports* 2022 12:1, vol. 12, no. 1, pp. 1–11, Feb. 2022, doi: 10.1038/s41598-022-05146-6.
- [9] H. Taherdoost, "Non-Fungible Tokens (NFT): A Systematic Review," *Information* 2023, Vol. 14, Page 26, vol. 14, no. 1, p. 26, Dec. 2022, doi: 10.3390/INFO14010026.
- [10] A. Kalimančević, M. Simić, and M. Radenković, "Blockchain Infrastructure for Managing Competences in Higher Education," in *Unlocking the Hidden Potentials of Organization Through Merging of Humans and Digitals, 19th International Symposium SymOrg*, 2024, pp. 86–93.
- [11] H. Li and D. Han, "EduRSS: A Blockchain-Based Educational Records Secure Storage and Sharing Scheme," *IEEE Access*, vol. 7, pp. 179273–179289, 2019, doi: 10.1109/ACCESS.2019.2956157.



- [12] C.-H. Wu and C.-Y. Liu, "Educational Applications of Non-Fungible Token (NFT)," *Sustainability* 2023, Vol. 15, Page 7, vol. 15, no. 1, p. 7, Dec. 2022, doi: 10.3390/SU15010007.
- [13] T. Sutikno and A. I. B. Aisyahrani, "Non-fungible tokens, decentralized autonomous organizations, Web 3.0, and the metaverse in education: From university to metaversity," *Journal of Education and Learning (EduLearn)*, vol. 17, no. 1, pp. 1–15, Feb. 2023, doi: 10.11591/EDULEARN.V17I1.20657.
- [14] W. Rehman, H. E. Zainab, N. Bawany, J. Imran, and N. Zakaria Bawany, "NFTs: *Applications and Challenges*," doi: 10.1109/ACIT53391.2021.9677260.
- [15] "Documentation - OpenZeppelin Docs." Accessed: Mar. 19, 2025. [Online]. Available: <https://docs.openzeppelin.com/>



CHALLENGING DEEPSEEK-R1 WITH SERBIAN HIGH SCHOOL MATH COMPETITION PROBLEMS

Nemanja Vučićević,
[0000-0002-4903-7280]

Marina Svičević*,
[0000-0003-2791-3849]

Aleksandar Milenković
[0000-0001-6699-8772]

University of Kragujevac,
Faculty of Science,
Kragujevac, Serbia

Abstract:

This paper investigates the application of the large language model DeepSeek-R1 in solving problems from the Serbian National Mathematics Competition for high school students (2023/2024), with the aim of examining its ability to understand, reason, and generate accurate mathematical solutions. The analysis covers all grade levels and both competition categories (A and B), comprising a total of 36 problems across various mathematical domains, including algebra, geometry, combinatorics, number theory, and logic. All problems were presented to the model in their original textual form, using LaTeX syntax to ensure accurate representation of mathematical expressions.

The results obtained by the model were compared to official grading criteria and to the average scores achieved by student participants. In several cases, DeepSeek-R1 reached scores that would have qualified for official awards, especially in the higher grades of category B. The model showed stronger performance on algebraic problems and those with a more formal structure, while it encountered difficulties with logic-based problems and less standard formulations. Most errors were due to occasional misinterpretation of the problem statements or the omission of key reasoning steps.

This research provides a realistic assessment of the capabilities of a contemporary large language model in solving complex mathematical problems. It also highlights possible directions for its integration as a support tool in the teaching and learning of mathematics.

Keywords:

Non-standard problem solving, DeepSeek-R1, Large Language Models, Math competitions.

INTRODUCTION

Mathematics competitions have a long-standing tradition in the Serbian education system and represent one of the most important forms of nurturing giftedness, developing logical thinking, and encouraging creative problem solving among students. From the lower grades of primary school to the end of secondary education, students have the opportunity to compete in solving demanding problems that often go far beyond the standard curriculum. These competitions not only recognize and affirm talent but also prepare students for future academic and professional achievements in fields such as mathematics, computer science, physics, and engineering.

Correspondence:

Marina Svičević

e-mail:

marina.svicevic@pmf.kg.ac.rs





One of the most prestigious events in this system is the National Mathematics Competition for High School Students, which represents the final stage following the municipal and district competitions. Participation at this level is limited to the most successful students, which gives the event particular weight in both educational and societal contexts. The problems featured at this stage are known for their difficulty—not merely testing mastered mathematical techniques, but encouraging students to independently recognize connections, construct solution strategies, and apply non-standard ideas. This demand for a creative approach to problem solving makes the problems especially challenging and valuable for analysis. The competition is organized in accordance with the Official Rulebook on Mathematics Competitions for Secondary School Students [1], which defines participation criteria, problem structure, and ranking methodology. The competition includes four grade levels (I–IV) and two categories: Category A, for students from the Mathematical Grammar School, and Category B, for students from all other grammar schools. In both categories, the exam lasts four hours, but the format differs: Category A has four problems worth up to 25 points each, while Category B has five problems worth up to 20 points, both totaling 100 points. Problems are adapted to the students' grade level and category, and cover areas defined by the Competition Curriculum [2], including logic, algebra, geometry, combinatorics, and number theory.

The increasing capability of large language models (LLM) to process and solve tasks involving formal reasoning raises important questions about their applicability in educational contexts and their potential use as tools for supporting learning or assessment. In this light, it is particularly interesting to examine how a modern model such as DeepSeek-R1 [3], which is not exclusively specialized in mathematics but demonstrates strong performance on logical reasoning tasks, performs when faced with problems of high complexity.

The main objective of this paper is to examine how the DeepSeek-R1 model would perform on problems from the Serbian National Mathematics Competition for high school students in the 2023/2024 school year, across both categories—A and B. By analyzing the solutions generated by the model, the study aims to assess its ability to understand problem statements, plan solutions, and produce precise mathematical reasoning in the context of problems designed for the most successful high school students in Serbia.

Within this objective, the paper addresses the following research questions:

- How many points would DeepSeek-R1 score if evaluated according to the official competition criteria by a human grading committee?
- Where would the model rank compared to real student participants in the same year?
- What are the characteristics of the problems where the solutions are incorrect, in terms of the mathematical domains they belong to, the way problems are formulated, and how the data provided in the problem statements are interpreted?

In addition to providing an empirical insight into the capabilities of an advanced artificial intelligence model, this study also contributes to the broader discussion of how LLMs can be integrated into educational practice—as tools for practice, diagnosis, or inspiration in the teaching of mathematics.

2. BACKGROUND AND RELATED WORK

In recent years, we have witnessed significant progress in the development of LLMs, which are increasingly capable of solving tasks well beyond the scope of natural language processing. One area that has attracted particular attention in recent literature is the application of these models in mathematics, where solving textual and symbolic problems—requiring reasoning, formal precision, and solution planning—has become a realistic possibility.

One of the first notable models focused on mathematical reasoning was Minerva, developed by Google and introduced in 2022 [4]. Trained on technical and scientific literature, including large corpora from ArXiv, Minerva achieved impressive results on the MATH dataset, with its 62-billion-parameter version reaching the average score on the Polish national high school graduation exam. The model employs a chain-of-thought technique, which enables gradual solution development through step-by-step explanations. GPT-4 [5] currently represents one of the most powerful available solutions, capable of solving problems at the Olympiad level. According to OpenAI's technical report, GPT-4 scores within the top 10% on standardized tests such as the SAT, GRE, and other math-related exams. A key strength of GPT-4 is its ability to generate logically coherent solutions with detailed reasoning steps, which significantly improves accuracy compared to previous model generations. A more recent contribution to this



domain is DeepSeekMath [6], which continued training on a corpus of over 120 billion tokens of mathematical content. Although relatively small (with 7 billion parameters), the model achieved 51.7% accuracy on the MATH dataset, and the use of self-consistency techniques increased its performance to 60.9%. Additionally, it incorporates Group Relative Policy Optimization (GRPO) to improve reasoning efficiency with limited resources. A notable advancement in prompt design is the MathPrompter technique [7], which combines multiple independent solution strategies for self-verification. This approach demonstrated that the accuracy of GPT-3.5 and GPT-4 on the MultiArith dataset of elementary arithmetic problems can be improved from 78.7% to 92.5% by employing parallel methods such as algebraic solutions, Python code, and verbal explanations.

In addition to individual performance analyses, there is a growing body of research focused on direct comparisons of different LLMs in solving mathematical problems of varying difficulty. Such comparisons allow for a more precise assessment of each model's strengths and limitations in terms of accuracy, robustness, sensitivity to formulation changes, and domain-specific reasoning. Recent evaluations have placed special emphasis on the DeepSeek-R1 [8] [9], OpenAI o1 [10], and the most recent, improved version o3-mini [11], all of which represent state-of-the-art LLMs optimized for reasoning tasks. DeepSeek-R1, based on a Mixture-of-Experts architecture and trained using reinforcement learning with a focus on mathematical reasoning, achieves outstanding results on the MATH-500 benchmark, reaching an accuracy of 97.3%, thereby outperforming the OpenAI o1 model, which scores 96.4% on the same dataset. These results were obtained using the self-consistency technique, which has become a standard in evaluating mathematical reasoning in modern LLMs [8]. Beyond its high accuracy, DeepSeek-R1 also demonstrates greater robustness to lexical and structural reformulations of problem statements compared to the o1 model, whose performance tends to degrade under such variations. Analyses show that this robustness is particularly evident in areas such as algebra and number theory, while more complex spatial problems, such as geometry, remain a challenge for all models. A comparison on the AIME 2024 benchmark confirms this close performance: DeepSeek-R1 solves 79.8% of the problems, while OpenAI o1 solves 79.2%, with the smaller o1-mini model falling significantly behind. In more recent evaluations, the o3-mini model, released in early 2025, shows even better performance on cat-

egories such as Olympiad-level and scientific problems, surpassing both previous architectures on the GPQA-Diamond dataset and the most difficult AIME tasks [11]. On the GSM8K dataset, which consists of elementary-level arithmetic word problems, GPT-4 achieves an accuracy of 92.0% [12], while DeepSeek-R1, using a self-consistent configuration, reaches 96.3% [13]. These results lead to a high level of precision in basic arithmetic tasks. In summary, comparisons among the most advanced LLMs show that models such as DeepSeek-R1 and OpenAI o3-mini have reached a level of mathematical reasoning comparable to that of top-performing human competitors across various levels. The differences between them are increasingly reflected in their robustness, domain-specific precision, and adaptability to reformulated problems, which represent key directions for future development in this field.

The standard evaluation of LLMs is mostly based on automatically comparing the generated answer with the expected solution, where accuracy is measured as the percentage of correctly solved problems. Although the problems are typically presented in textual form, solving them requires multi-step logical reasoning, which has led to the adoption of more advanced evaluation methods such as self-consistency, partial credit scoring, and the assessment of intermediate steps. Chen et al. [14] demonstrated that GPT-4o [15] is capable of assigning partial credit to student-generated solutions with a high level of agreement with human graders (70–80%), using multiple evaluation passes and a detailed scoring rubric.

While modern models have achieved remarkable results on existing problem sets, there remains a need for further evaluation in new contexts—especially on problems that were not part of their training data. Such cases—like problems from national mathematics competitions—offer valuable insight into the true capabilities of these models and the possibilities for their application in educational and competitive settings, while also presenting an added challenge in evaluating the quality of generated solutions.



3. METHODOLOGY

This section of the paper outlines the steps taken to examine the ability of the large language model DeepSeek-R1 to solve mathematics problems from the Serbian National Mathematics Competition for high school students in the 2023/2024 school year. The choice of both the model and the problem set is based on two key criteria: on the one hand, DeepSeek-R1 is one of the most advanced open-source models specifically trained for mathematical reasoning; on the other hand, the competition problems represent an authentic and challenging dataset that spans a wide range of mathematical topics, formulation styles, and diverse solution approaches.

This setup makes it possible to test the model not only on well-known and frequently used datasets but also on real-world problems that were not part of its training data, offering valuable insight into its ability to generalize and apply reasoning in an educational context.

3.1. DATASET: SERBIAN NATIONAL MATHEMATICS COMPETITION 2023/2024

The dataset used in this study consists of original problems from the Serbian National Mathematics Competition for high school students, held during the 2023/2024 school year. The competition is organized by the Mathematical Society of Serbia, and the problems span a wide range of high school mathematics topics, including logic, algebra, geometry, combinatorics, number theory, and elements of analysis.

The competition is divided into four grade levels and two categories:

- Category A is intended for students from mathematical grammar schools,
- Category B is intended for students from all other grammar schools.

For each grade, Category A includes four problems, while Category B includes five. Problems in Category A are worth 25 points each, and those in Category B are worth 20 points each, giving a maximum of 100 points in both cases. In the 2024 competition, all problems were presented in textual form, without diagrams, illustrations, or graphical aids. The formulations typically require multiple steps and abstract thinking, making them well-suited for testing LLMs in the domain of mathematical reasoning. Since modern models support LaTeX syntax, the problem statements in this study were

formatted accordingly, allowing for accurate representation of mathematical expressions and symbols. Each problem was presented to the model as an independent prompt in Serbian, without additional context, in order to simulate realistic solving conditions. The evaluation of the responses was performed manually by a current member of the official competition committee, based on the official solutions and the same criteria used during the actual competition. Special attention was given to the accuracy of the final answer, the logical coherence of the reasoning steps, and the identification of typical errors.

3.2. DEEPSEEK-R1

The DeepSeek-R1 model [16] is a next-generation large language architecture developed by DeepSeek AI with the goal of enhancing models' abilities to perform complex reasoning, particularly in the domains of mathematics and programming. Released as an open-source model in early 2025, DeepSeek-R1 is based on a Mixture-of-Experts (MoE) architecture, with a total capacity of 671 billion parameters, of which approximately 37 billion are active per query. This approach enables a better balance between response quality and execution efficiency.

Unlike some earlier LLMs that relied primarily on supervised learning, DeepSeek-R1 was trained using reinforcement learning, where the model was optimized directly based on the quality of its logical reasoning. Specifically, the training employed a “reasoning-in-the-loop” technique, allowing the model to learn from chains of thought and feedback on the correctness of its conclusions. The training process included a strong emphasis on the mathematical domain, further reinforced through fine-tuning on a large collection of problems, including datasets such as MATH, GSM8K, AIME, and OlympiadBench.

DeepSeek-R1 supports an input context of up to 128,000 tokens, making it well suited for handling long mathematical formulations and complex problem statements. In addition to the full model, the team has released several distilled versions, including R1-Zero, a variant without additional fine-tuning, which contributes to transparency in research settings.

The choice of DeepSeek-R1 for this study is based on several reasons. First, it is open source, which allows for direct access and reproducibility of results. Second, prior experiments have shown that DeepSeek-R1 performs on par with, and often surpasses, closed models such as



OpenAI o1, particularly on mathematical benchmarks. Third, its training is explicitly focused on reasoning, making it highly relevant for the types of tasks featured in mathematics competitions, where modeling multi-step solutions is more critical than simply generating a correct final answer.

4. RESULTS AND DISCUSSION

This section presents the results of applying the DeepSeek-R1 model to the problems from the Serbian National Mathematics Competition 2023/2024. The analysis covers a total of 36 problems, four in Category A and five in Category B for each of the four grade levels. In each case, the model was given only the textual formulation of the problem, presented in LaTeX format, without additional instructions, examples, or context, and consistent with the original tasks, which did not include any illustrations. The answers generated by the model were manually reviewed according to the official scoring criteria. The results obtained were compared with the average scores achieved by human participants by grade and category. Additionally, an analysis was conducted to determine which types of problems the model solved more or less successfully. Special attention was given to the qualitative analysis of errors and to the model’s potential applications in educational contexts.

The first step in the evaluation involved comparing the total number of points that DeepSeek-R1 achieved per grade with the average number of points earned by students in the same categories at the 2023/2024 National Mathematics Competition.

The results are shown separately for Category A, which includes students from mathematical grammar schools, and Category B, which includes students from all other grammar schools (Figure 1). In both categories, the model scored below average in the first grade but generally outperformed the average results of students in the higher grades. Although there is no clear upward trend across all grades, the model shows consistent advantages in the second, third, and fourth grades, especially in the second grade of both categories. It is important to note that neither problem difficulty nor grade level alone necessarily explain the model’s performance. According to the grading committee, problems for the first grade often include non-standard formulations, with an increased presence of combinatorial and logic-oriented tasks, which may have contributed to the model’s lower performance at this level. Models like DeepSeek-R1 tend to be more effective at solving problems with a more formal structure, which aligns with the stronger results observed in grades dominated by algebraic and analytical problems. Overall, the diagram confirms that the model performs significantly better in Category B, where the problems are generally less complex.

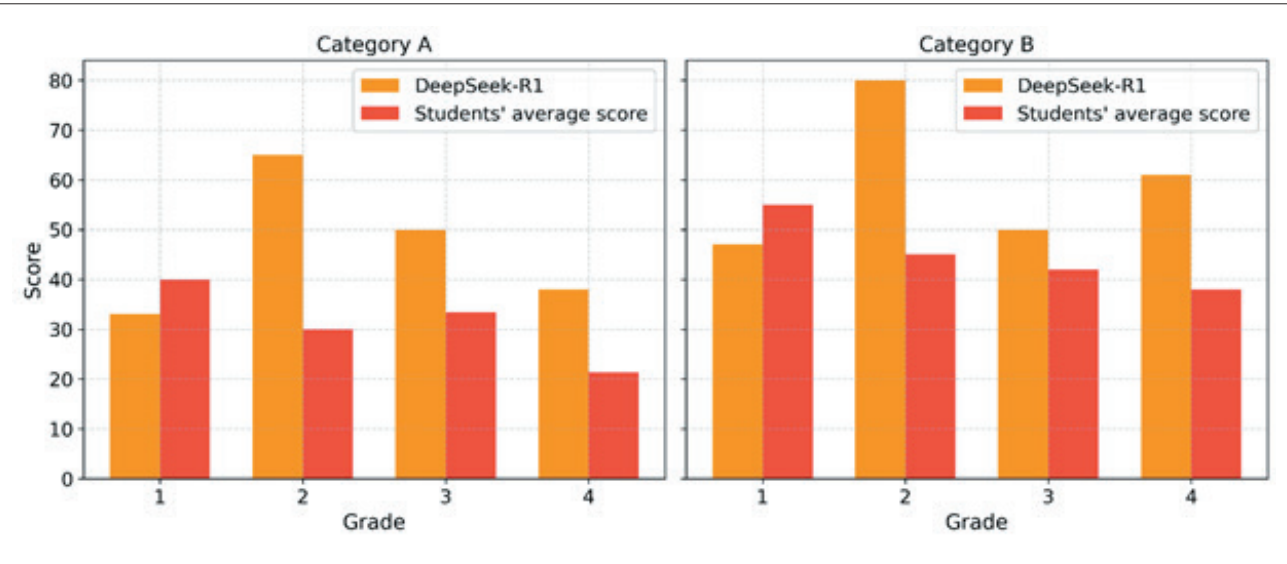


Figure 1. Comparison of DeepSeek-R1's results with the students' average scores achieved by students at the National Mathematics Competition 2023/2024, across both A and B categories



In addition to comparing DeepSeek-R1's total score with average values, it is also insightful to examine the model's placement on the official competition ranking lists by grade and category, including visual indicators of awarded prizes (Figure 2). In the first grade, the model did not manage to reach an award-winning position in either category — it ranked 33rd out of 50 in Category A and 31st out of 42 in Category B. However, from the second grade onward, the model achieved significantly better results. In Category A, it placed 4th out of 41 in the second grade, earning second prize, and maintained a top-ten placement in the third and fourth grades, earning third prizes. Even better results were achieved in Category B, where the model placed 3rd out of 67 in the second grade, earning first prize, and secured 13th and 7th place in the third and fourth grades, respectively, both corresponding to second prizes. These results indicate that the model is most successful in Category B from the second grade onward, which aligns with the relatively lower complexity of problems in that category. At the same time, the model's consistent placement in the top third of Category A in the higher grades demonstrates its capacity to handle more challenging problems as well.

The results show that the DeepSeek-R1 model, when evaluated according to official competition criteria, can achieve high rankings and even win awards in certain categories, especially in higher grades and with less demanding problems. Its errors most frequently occur in tasks with ambiguous formulations, logic-based

reasoning, or geometry problems that require constructive or proof-based solutions, indicating limitations in the model's ability to comprehend abstract problems and carry out multi-step reasoning. These insights confirm that LLMs have the potential to serve as valuable educational tools in mathematics instruction — for independent practice, reasoning diagnostics, and encouraging diverse problem-solving approaches.

5. CONCLUSION

This study has shown that the large language model DeepSeek-R1 is capable of successfully solving high-complexity mathematical problems, such as those featured in the National Mathematics Competition for Serbian high school students. Through systematic evaluation, it was found that the model not only achieves results comparable to top-performing contestants in certain categories but even reaches rankings that would earn it official awards in specific grades and categories. Its highest effectiveness was observed in algebraic and analytical problems, while weaker performance was noted in logic-based tasks and less clearly formulated problems.

These findings suggest that LLMs like DeepSeek-R1 have already reached a level that makes them relevant in educational contexts—not as replacements for students or teachers, but as complementary tools for practice, automated assessment, and the encouragement of creative problem-solving. The unique value of such analyses lies

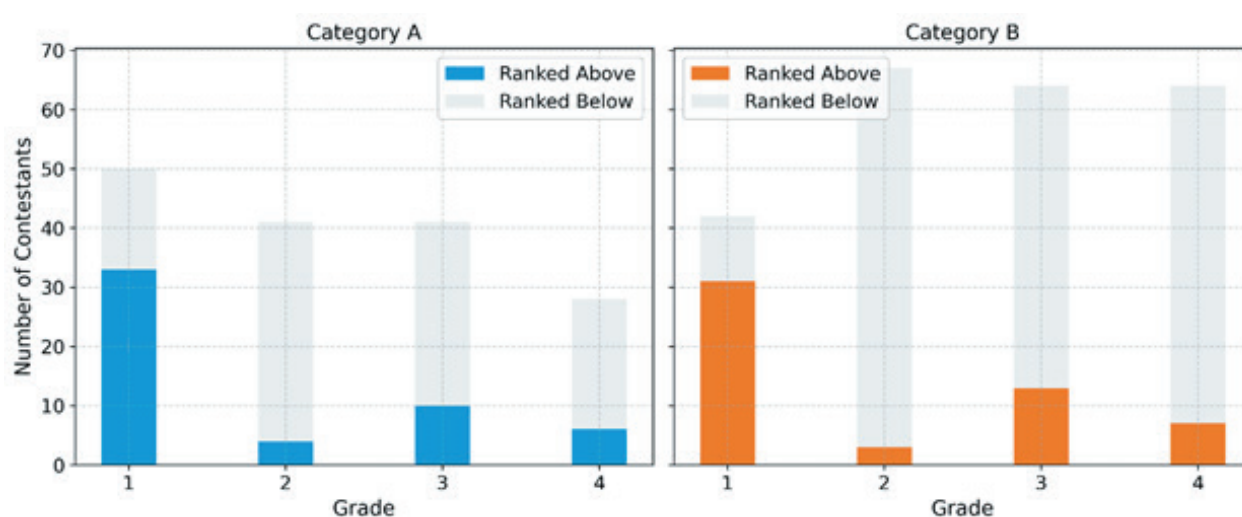


Figure 2. DeepSeek-R1's relative ranking among contestants at the National Mathematics Competition 2023/2024, across both Category A and Category B. Each bar shows the number of students ranked above the model (colored) and those ranked below (light gray)



in the fact that competition problems represent challenging, real-world tasks that are not part of the models' training datasets, thus offering a more objective view of their capabilities and limitations.

Future research could involve comparative evaluations of multiple models on the same problem set, as well as an exploration of how LLMs might be integrated into mathematics education through interactive platforms and tailored feedback mechanisms.

6. ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia through the Agreement No. 451-03-65/2024-03/200122.

REFERENCES

- [1] Društvo matematičara Srbije, "Pravilnik o takmičenjima," [Online]. Available: https://dms.rs/wp-content/uploads/2021/12/Pravilnik_o_takmičenjima_SS_matematika.pdf.
- [2] Društvo matematičara Srbije, "Program takmičenja," [Online]. Available: https://dms.rs/wp-content/uploads/2016/12/program_2016.pdf.
- [3] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang and X. Bi, "Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.
- [4] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag and T. Gutman-Solo, "Solving quantitative reasoning problems with language models," *Advances in Neural Information Processing Systems*, vol. 35, p. 3843–3857, 2022.
- [5] OpenAI, "GPT-4," [Online]. Available: <https://openai.com/index/gpt-4-research/>. [Accessed 10 March 2025].
- [6] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li and Y. Wu, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.
- [7] S. Imani, L. Du and H. Shrivastava, "Mathprompter: Mathematical reasoning using large language models," *arXiv preprint arXiv:2303.05398*, 2023.
- [8] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang and X. Bi, "Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.
- [9] DeepSeek. [Online]. Available: <https://www.deepseek.com/>. [Accessed 10 March 2025].
- [10] OpenAI, "o1," [Online]. Available: <https://openai.com/index/openai-o1-system-card/>. [Accessed 10 March 2025].
- [11] OpenAI, "o3-mini," [Online]. Available: <https://openai.com/index/openai-o3-mini/>. [Accessed 10 March 2025].
- [12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman and S. Anadkat, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.
- [13] A. Eassa, A. Shah, H. Mao, H. Lu, E. Ho, J. Xin and O. Almog, "NVIDIA," [Online]. Available: <https://developer.nvidia.com/blog/nvidia-blackwell-delivers-world-record-deepseek-r1-inference-performance/>. [Accessed 20 March 2025].
- [14] Z. Chen and T. Wan, "Using Large Language Models to Assign Partial Credits to Students' Problem-Solving Process: Grade at Human Level Accuracy with Grading Confidence Index and Personalized Student-facing Feedback," *arXiv preprint arXiv:2412.06910*, 2024.
- [15] OpenAI, "GPT-4o," 13 May 2024. [Online]. Available: <https://openai.com/index/gpt-4o-system-card/>. [Accessed 10 March 2025].
- [16] DeepSeek-R1, 2025. [Online]. Available: <https://huggingface.co/deepseek-ai/DeepSeek-R1>.



THE DESIGN CHARACTERISTICS OF INTELLIGENT TUTORING SYSTEMS FOR STEM EDUCATION

Veljko Aleksić^{1*},
[0000-0003-2337-1288]

Dionysios Politis²
[0009-0005-2876-7283]

¹University of Kragujevac,
Faculty of Technical Sciences in Čačak,
Čačak, Serbia

²Aristotle University of Thessaloniki,
Faculty of Sciences,
Thessaloniki, Greece

Abstract:

This paper comprehensively examines intelligent tutoring systems as transformative educational technology that leverages artificial intelligence in creating autonomous adaptive digital learning environments for STEM education. The research articulates a sophisticated four-component framework design for delivering personalized instruction aligned with pedagogical principles. We analyzed advanced probabilistic approaches that enable the dynamic adaptation of learning pathways, content sequencing, and difficulty calibration based on continuous assessment of student knowledge states. Our investigation was extended to personalized feedback mechanisms that monitor problem-solving processes, identify misconceptions, and provide contextual guidance through natural language processing and affective computing techniques. The empirical evidence from diverse STEM disciplines demonstrated that well-designed intelligent tutoring systems significantly outperform traditional instructional methods regarding learning outcomes, knowledge retention, and student engagement. Through a detailed case analysis of exemplary systems, we identified critical design characteristics that contribute to educational effectiveness. The presented findings have significant implications for educational policy, curriculum design, and the development of next-generation intelligent tutoring systems that can effectively address the complex, interdisciplinary nature of contemporary STEM education.

Keywords:

Intelligent Tutoring Systems, STEM, Education.

INTRODUCTION

Personalized learning educational approaches supported by artificial intelligence (i.e., AI) techniques and technology have the potential to address the diverse needs of STEM students (i.e., Science, Technology, Engineering, and Mathematics – STEM), taking into account the interdisciplinary and transdisciplinary nature of educational contents as well as the capabilities and characteristics of individual learners [1]. Intelligent tutoring system (i.e. ITS) represents an educational technology utilized for providing targeted feedback, advice, guidance, and explanations to improve knowledge acquisition, enhance conceptual understanding, develop practical skills, and strengthen the competencies of individual students.

Correspondence:

Veljko Aleksić

e-mail:

veljko.aleksic@ftn.kg.ac.rs





ITS adaptive characteristics facilitate the dynamic adjustment of difficulty levels and complexity of tasks and materials, ensuring students are intellectually challenged at appropriate personalized levels [2]. The implementation of personalized learning in STEM education via ITS demonstrated significant potential in enhancing the degree of learning outcome achievement, increasing student engagement, and fostering self-regulated learning behaviors [3]. By providing students with customized support and guidance, ITS can effectively bridge knowledge gaps, address conceptual misconceptions, and promote a deeper understanding of the learned material. However, the effectiveness of implementing such systems in educational contexts critically depends on the quality of underlying models, the accuracy of established student profiles, the robustness of algorithms, and alignment with scientifically validated and accepted pedagogical principles [4].

2. INTELLIGENT TUTORING SYSTEMS

Intelligent tutoring systems are a significant educational technology advancement compared to learning management systems, as they utilize machine learning algorithms and AI to provide students with an autonomous, adaptive, and personalized learning environment. Through the application of AI techniques, these systems can dynamically assess students' specific knowledge and skills, provide individualized feedback, and optimize learning pathways to meet the needs of each individual [5]. One of the key advantages is the capacity to continuously assess students' knowledge and adjust instructional content accordingly. By employing machine learning algorithms and educational data mining techniques, these systems can analyze student interactions, responses, and performance patterns to create accurate models of their knowledge and skills. This assessment enables the system to identify gaps in understanding, misconceptions, and areas requiring additional support in real time [6]. Unlike traditional computer-based instruction that relies on pre-programmed feedback, ITS can generate dynamic and contextually relevant feedback tailored to students' specific needs. This feedback can take various forms, such as hints, explanations, examples, and guidelines, which are adapted to the student's current level of understanding and learning style [7]. In addition, ITS can dynamically adjust learning pathways and the sequence of instructional content based on student performance and progress. Through continuous monitoring of student interactions and adaptation of difficulty

levels, pace, and the scope of educational materials, personalized learning trajectories can be created to optimize the learning experience. This adaptive sequencing ensures that students are presented with balanced challenges to maintain optimal levels of engagement and motivation [8].

2.1. GOOD PRACTICE EXAMPLES

Research findings demonstrate that the implementation of ITS as educational technology is effective in improving the achievement level of learning outcomes [9], and that they outperform traditional instruction and other derived forms of computer-based instruction (e.g., blended learning) in terms of learning progress and efficiency [10]. These systems show significant positive effects on students' academic achievement across various domains, including STEM [11]. Educational technology integration into various school subjects opened new possibilities for personalized and adaptive learning experiences. In the domain of physics education, the *Andes* ITS provides an interactive learning environment where students solve problems and receive immediate feedback and further guidance. *Andes* uses a Bayesian network to model students' knowledge states and adapts feedback and problem selection based on their individual needs [12]. The *Rimac* is a sophisticated adaptive tutoring platform designed to address persistent challenges in physics education through the implementation of knowledge construction dialogues (i.e., KCDs) [13] integrating an advanced student modeling component that dynamically assesses knowledge based on pretest responses and dialogue interactions. The *ORCCA* intelligent tutor leverages the *CTAT* [14] rule engine to deliver an adaptive chemistry homework experience through a paper-like free-form workspace coupled with dynamic feedback mechanisms [15]. This system provides students with personalized guidance during problem-solving while offering teachers valuable insights into learning challenges, representing a significant advancement over traditional digital assessment methods in chemistry education. *MetaTutor* can be used as a hypermedia-based ITS for learning biology that employs pedagogical agents to deliver adaptive scaffolding through strategic prompts and feedback in the experimental condition while allowing unrestricted exploration without guidance in the control condition [16]. Its technical architecture integrates three resource categories: content materials, experimental protocol parameters, and condition-specific workflows. *StuDiAsE* is an ITS that leverages AI to assess comprehension, evaluate prior knowledge, and deliver personalized educational



assistance based on individual learner profiles [17]. The system integrates five core subsystems (monitoring, logging, profiling, modeling, and evaluation) that work cohesively to track and enhance the learning experience in engineering education. The platform provides differentiated interfaces that allow learners to navigate educational materials with adaptive guidance while enabling educators to modify content and assess learning outcomes. *TECH8* is an effective individualized tutoring system for learning mechanical engineering, with the experimental group achieving 55.7% on summative assessments compared to 44.3% in the traditional teaching control group [18]. Its implementation resulted in measurable cognitive advancement, with 23.7% of students progressing to higher cognitive levels. *Active Math* ITS dynamically creates learning materials based on individual student preferences, knowledge levels, and learning objectives by providing interactive exercises, explanations, and examples tailored to the assessed individual needs of students, thus indicating very high efficacy in promoting self-regulated learning and enhancing mathematical problem-solving skills [19]. Based on the *Cognitive Tutor* ITS that has been successfully implemented to provide guidance and feedback to students for complex mathematical problem-solving [20], the *MATHia* ITS was developed for individual mathematics formative and summative assessment and successfully used in primary and secondary education [21]. Currently, the most popular mathematics blended learning intelligent tutoring platform is *ALEKS*. The system presents students with an individualized sequence of questions guiding them through the problem-solving process, consequently requiring very little teacher involvement in the learning process [22]. The advancements in natural language processing (i.e., NLP) and dialogue systems have enabled the development and modernization of conversational ITS that now can engage students in direct interaction using natural language, understand students' questions, provide explanations, and adaptively guide them through the problem-solving process using natural language vocal conversation. This approach enhances students' conceptual understanding and problem-solving abilities in target domains. By simulating the interactive nature of human teaching, conversational ITS creates immersive learning experiences [23] [24]. The development of ITS requires significant investment in terms of expertise, resources, and time. Creating accurate domain models, student models, and pedagogical strategies relies on close collaboration between subject teachers, instructional designers, and researchers in the field of programming, machine learning, and AI.

3. DESIGN CHARACTERISTICS

The ITS architecture constitutes a sophisticated framework comprising four principal components:

1. **Domain Model** – represents the knowledge and skills that the ITS teaches. It encompasses domain-specific expertise, including concepts, facts, procedures, and problem-solving strategies. The domain model serves as the foundation for generating instructional content, assessment elements, and formulating feedback [25]. Various knowledge representation techniques are employed to structure domain knowledge in digitally readable formats, such as ontologies, semantic networks, and rules. For instance, the domain model of mathematics education ITS incorporates mathematical concepts, theorems, problem-solving strategies, and common misconceptions. The domain model facilitates the generation of contextualized explanations and examples based on learner interactions with the application. It enables the system to evaluate student problem-solving steps, identify errors or misconceptions, and provide appropriate feedback to guide the learner toward successful mastery of the material [26].
2. **Student Model** – constitutes a critical component of ITS in education as it records current knowledge, preferences, and other relevant learner characteristics. This model represents a dynamic representation of the student's understanding and progress within the domain, enabling the system to adequately adapt instructional strategies and content delivery [27]. The student model is continuously updated based on learner interactions with the system (e.g., responses to questions, problem-solving attempts, and navigation choices). To establish a model of student knowledge, techniques such as overlay modeling, perturbation modelling, and knowledge tracing are utilized. For example, language learning ITS tracks learners' proficiency levels across different linguistic skills (grammar, vocabulary, pronunciation) and adjusts the difficulty level and content of lessons accordingly [28]. With an accurate representation of student strengths and weaknesses, the ITS can provide valid personalized recommendations, exercises, and feedback.



3. **Tutoring Model** – is designed for pedagogical decision-making and implementing instructional strategies employed by the ITS. This model determines how the system interacts with the learner by selecting appropriate instructional actions based on the student's knowledge level, learning objectives, and interaction context. The tutoring model employs various pedagogical theories, strategies, and instructional design principles to optimize the learning experience, such as adaptive feedback techniques, providing hints, and selecting appropriate problems to monitor student progress and promote understanding of the learned material. The model also incorporates techniques from the domains of teaching methodology and cognitive psychology to enhance instructional efficiency, e.g., content repetition, interleaving, and retrieval practice aimed at effective knowledge transfer and long-term retention of learned material [29].
4. **User Interface** – represents the communication channel between the learner and the ITS that manages learning content, processes student inputs, and generates feedback and guidance. The user interface design should be intuitive, interactive, and visually appealing to enhance student motivation and engagement [30]. Navigation should be clear, the layout consistent, and content comprehensible and accessible to easily accommodate diverse student needs and preferences [31]. For example, the user interface of programming learning ITS must possess functions such as a code editor, debugger, and visualization tools. Feedback on potential syntax errors or logical inconsistencies must be provided in real time. The ultimate goal of a proficient user interface is to facilitate seamless interaction between the learner and the ITS, preferably enabling natural language dialogue, multimodal input, and adaptive content presentation. Advances in NLP and dialogue systems have enabled the development of conversational systems that can engage learners in interactive discussions, answer questions, and provide explanations [32].

Through the application of AI techniques and educational theories, ITS revolutionized the approach to teaching and learning in the digital age. However, their successful implementation and further development necessitate a further multidisciplinary approach, involving experts from education, psychology, and computer science,

to ensure that the systems are pedagogically valid, user-friendly, and effective in achieving desired learning outcomes.

3.1. ADAPTIVE LEARNING PATHWAYS

Adaptive learning pathways constitute a pivotal characteristic of ITS that facilitates personalized instruction tailored to individual student needs. Through continuous assessment of student knowledge and performance, the system dynamically adjusts the sequence, difficulty, and pace of learning content to create an optimal learning experience. Adaptivity is fundamental in educational applications as it ensures students receive content appropriate to their current level of understanding and promotes efficient learning [33]. In the development of adaptive learning pathways, Bayesian Knowledge Tracing (i.e. BKT) is predominantly employed as a probabilistic model that evaluates student knowledge based on their performance on tasks and assessments [34]. This model represents student knowledge as a set of binary variables, indicating, for example, whether a skill has been mastered or not. The probability of skill mastery is updated following each student interaction, taking into account factors such as response accuracy and number of attempts. By inferring student knowledge states, BKT enables ITS to adapt the selection and sequence of learning tasks to optimize student knowledge and skill acquisition [35]. An alternative to the aforementioned approach is Item Response Theory (i.e. IRT), a psychometric approach that models the relationship between students' ability and their responses to assessment items [36]. The model estimates students' latent abilities based on their performance on a set of items with known difficulty and discrimination parameters, thus enabling the ITS to select items that are most informative for assessing student abilities and to adjust content difficulty based on their expertise. This adaptivity ensures that students are appropriately challenged, preventing boredom or frustration [37]. Adaptive learning pathways are implemented through various algorithms and decision rules that determine the optimal sequence and pace of instruction for each student. These algorithms consider factors such as student prior knowledge, learning objectives, cognitive abilities, and affective states. For instance, Knowledge Space Theory (i.e. KST) is an approach that represents domain knowledge as a network of prerequisites and dependencies [38]. Intelligent tutoring systems based on this theory can generate personalized learning pathways by identifying the



most efficient route through the knowledge space based on the student's current knowledge and desired learning objectives [39]. Given that adaptive learning pathways are not limited solely to content selection and sequencing but also encompass the adaptation of instructional strategies and feedback, it is possible to utilize the expertise reversal effect, which suggests that instructional support effective for novices may be detrimental for more advanced students [40]. Similarly, adaptive feedback can be created based on student misconceptions, errors, and problem-solving strategies, offering personalized explanations and hints to promote understanding [41]. The efficacy of adaptive learning pathways in ITS has been demonstrated in various educational technology applications. For example, the *Cognitive Tutor Algebra* system, which utilizes BKT for skill modeling and adaptive task selection, has shown significant improvements in students' abilities to solve mathematical problems compared to traditional instruction [42]. Similarly, the *ALEKS* intelligent system employs KST to generate adaptive learning pathways, enhancing student achievement and engagement, also in mathematics [43]. One of the primary challenges in the broader ITS implementation is the necessity of defining precise and comprehensive domain models that encompass complex relationships between skills and concepts and achieving an appropriate balance between system control and student autonomy in adaptive learning pathways. Although the system can optimize the learning experience based on data-driven decisions, it is important to consider student preferences as well as self-regulated learning goals and strategies. Providing students with a degree of control over their learning pathways and enabling exploratory learning enhances student motivation and engagement [44].

3.2. PERSONALIZED FEEDBACK AND GUIDANCE

Personalized feedback is a critical component of each ITS designed to support the development of problem-solving skills and foster a deeper understanding of instructional content. By continuously monitoring student problem-solving steps and their comparison with expert models or predefined solution pathways, the system can identify errors, misconceptions, or suboptimal strategies. This real-time analysis enables the generation of immediate feedback that highlights specific errors or challenges faced by the student. For example, programming education ITS can identify and indicate syntax errors, logical inconsistencies, or inefficient code structures in real time, guiding the student toward cor-

rect solutions [45]. ITS employs various techniques to provide personalized instructions and explanations that support students' problem-solving processes. When a student encounters difficulty while solving a specific task, the system can display steps that gradually reveal information or suggest problem-solving strategies. This scaffolded learning approach encourages students to think critically and arrive at solutions independently [16]. Additionally, the ITS can generate explanations of specific concepts and provide justifications for certain problem-solving choices or connect the current problem with previously acquired knowledge and skills. Such targeted elaborations help students develop a deeper understanding of the instructional content and enhance their metacognitive abilities [46]. NLP techniques have significantly enhanced the interactive and conversational capabilities of ITS, enabling it to "understand" and interpret student inputs (e.g., questions and explanations) in natural human language. This simulation of human tutoring conversation facilitates more natural interaction. Beyond providing feedback on problem-solving steps and conceptual understanding, ITS can adapt its feedback based on students' affective states. Affective computing allows the system to detect and respond to students' emotional states, such as frustration, confusion, or boredom, which can significantly impact learning outcomes and motivation. By analyzing students' facial expressions, eye movements, or various physiological signals, the ITS can infer their affective states and accordingly adjust the feedback. For instance, if a student appears frustrated or disinterested, the system can provide encouraging feedback, offer additional support, or suggest alternative learning strategies to maintain motivation and promote persistence [47]. Designing effective feedback and guidance in ITS requires achieving a balance between providing sufficient support and promoting student autonomy; specifically, the timing, specificity, and adaptivity of feedback must be carefully balanced to achieve learning outcomes [48]. Domain models that enable the generation of contextually relevant feedback must be precisely and comprehensively defined. For example, incorporating NLP capabilities into ITS feedback generation requires the use of advanced information technologies and linguistic knowledge, as human language is highly complex and often ambiguous.



4. CONCLUSION

Integrating ITS into existing educational systems and curricula requires careful planning, teacher training, and continuous support to ensure their effective adoption and utilization. To address these challenges and further advance the ITS field, a new direction of development involves the integration of data mining techniques and learning analytics to improve student modeling and adaptation. By utilizing the large amounts of educational data generated by ITS, data mining algorithms can reveal patterns and details that enable more precise student modeling. Learning analytics are used in visualizing and interpreting learning patterns, facilitating data-driven decision-making by teachers and researchers, and optimizing the learning experience [49]. Another alternative research direction involves incorporating Open Learner Models (i.e., OLMs) into ITS to promote metacognition and self-regulated learning. This approach provides students with access to their models, allowing them to review results, and reflect on their identified knowledge, progress, and learning strategies. By making the student model transparent and interactive, OLMs can foster self-awareness, goal-setting, and the development of self-monitoring skills. Research indicates that these models enhance student motivation, engagement, and achievement in learning environments supported by ITS [50]. A third direction is exploring the application of affective computing techniques to create “emotional” ITS, as they are focused on recognizing, interpreting, and generating emotions in human-computer interactions. This makes ITS able to detect and respond to students' emotional states, such as frustration, confusion, or boredom. Emotionally intelligent tutoring systems can provide personalized support and adaptations that consider students' affective needs, thereby improving their engagement, motivation, and learning outcome achievement [51].

REFERENCES

- [1] I. Celik, M. Dindar, H. Muukkonen and S. Järvelä, "The Promises and Challenges of Artificial Intelligence for Teachers: a Systematic Review of Research," *TechTrends*, vol. 66, no. 4, p. 616–630, 2022.
- [2] H. Khosravi, S. Sadiq and D. Gasevic, "Development and Adoption of an Adaptive Learning System," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, Portland, USA, 2020.
- [3] F. de Moraes and P. A. Jaques, "Improving Sensor-Free Affect Detection by Considering Students' Personality Traits," *IEEE Transactions on Learning Technologies*, vol. 17, p. 542–554, 2024.
- [4] V. Aleksić and D. Politis, "Trait Emotional Intelligence and Multiple Intelligences as Predictors of Academic Success in Serbian and Greek IT Students," *International Journal of Cognitive Research in Science, Engineering and Education (IJCRSEE)*, vol. 11, no. 2, p. 173–185, 2023.
- [5] J. D. Gobert, M. A. Sao Pedro, H. Li and C. Lott, "Intelligent tutoring systems: a history and an example of an ITS for science," in *International Encyclopedia of Education (Fourth Edition)*, Elsevier, 2023, p. 460–470.
- [6] V. Aleven, I. Roll, B. M. McLaren and K. R. Koedinger, "Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 1, p. 205–223, 2016.
- [7] A. Alam, "Harnessing the Power of AI to Create Intelligent Tutoring Systems for Enhanced Classroom Experience and Improved Learning Outcomes," in *Intelligent Communication Technologies and Virtual Mobile Networks, Singapore*, Springer Nature Singapore, 2023, p. 571–591.
- [8] S. Feng, A. J. Magana and D. Kao, "A Systematic Review of Literature on the Effectiveness of Intelligent Tutoring Systems in STEM," in *2021 IEEE Frontiers in Education Conference (FIE)*, Lincoln, USA, 2021.
- [9] T. Son, "Intelligent Tutoring Systems in Mathematics Education: A Systematic Literature Review Using the Substitution, Augmentation, Modification, Redefinition Model," *Computers*, vol. 13, no. 10, p. 270, 2024.
- [10] E. Mousavinasab, N. R. Zarifsanaiey, S. Niakan Kalhori, M. Rakhshan, L. Keikha and M. Ghazi Saeedi, "Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods," *Interactive Learning Environments*, vol. 29, no. 1, p. 142–163, 2018.
- [11] W. Xu and F. Ouyang, "The application of AI technologies in STEM education: a systematic review from 2011 to 2021," *International Journal of STEM Education*, vol. 9, no. 1, p. 59, 2022.
- [12] K. VanLehn, B. van de Sande, R. Shelby and S. Gershman, "The Andes Physics Tutoring System: An Experiment in Freedom," in *Advances in Intelligent Tutoring Systems*, Springer, 2010, p. 421–443.
- [13] T. Kodama, R. Tanaka and S. Kurohashi, "Construction of Hierarchical Structured Knowledge-based Recommendation Dialogue Dataset and Dialogue System," in *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, Dublin, Ireland, 2022.



- [14] V. Aleven, J. Sewall, O. Popescu, M. van Velsen, S. Demi and B. Leber, "Reflecting Chapter 22 on Twelve Years of ITS Authoring Tools Research with CTAT," in *Design Recommendations for Intelligent Tutoring Systems: Authoring Tools and Expert Modeling Techniques*, Springer, 2015, p. 263.
- [15] E. C. King, M. Benson, S. Raysor, T. A. Holme, J. Sewall, K. R. Koedinger, V. Aleven and D. J. Yaron, "The Open-Response Chemistry Cognitive Assistance Tutor System: Development and Implementation," *Journal of Chemical Education*, vol. 99, no. 2, p. 546–552, 2022.
- [16] R. Azevedo, F. Bouchet, M. Duffy, J. Harley, M. Taub, G. Trevors, E. Cloude, D. Dever, M. Wiedbusch, F. Wortha and R. Cerezo, "Lessons Learned and Future Directions of MetaTutor: Leveraging Multichannel Data to Scaffold Self-Regulated Learning With an Intelligent Tutoring System," *Frontiers in Psychology*, vol. 13, p. 813632, 2022.
- [17] M. Samarakou, E. Fylladitakis, W. G. Fruh, A. Hatzia Apostolou and M. Grigoriadou, "How eLearning Affects The Motivation Of Higher Education Students: A Case Study For StuDiAsE," in *Global Learn 2015*, Berlin, Germany, 2015.
- [18] K. Dolenc and B. Aberšek, "TECH8 intelligent and adaptive e-learning system: Integration into Technology and Science classrooms in lower secondary schools," *Computers & Education*, vol. 82, pp. 354–365, 2015.
- [19] G. Rebolledo-Mendez, N. S. Huerta-Pacheco, R. S. Baker and B. du Boulay, "Meta-Affective Behaviour within an Intelligent Tutoring System for Mathematics," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 1, p. 174–195, 2021.
- [20] I. Arroyo, B. P. Woolf, W. Burelson, K. Muldner, D. Rai and M. Tai, "A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect," *International Journal of Artificial Intelligence in Education*, vol. 24, no. 4, p. 387–426, 2014.
- [21] H. Almoubayyed, R. Bastoni, S. R. Berman, S. Galasso, M. Jensen, L. Lester, A. Murphy, M. Swartz, K. Weldon, S. E. Fancsali, J. Groppen and S. Ritter, "Rewriting Math Word Problems to Improve Learning Outcomes for Emerging Readers: A Randomized Field Trial in Carnegie Learning's MATHia. Artificial Intelligence in Education," in *International Conference on Artificial Intelligence in Education*, Cham: Springer Nature Switzerland, 2023.
- [22] A. Mangum and J. Sorrells, "Utilizing ALEKS and Standards Grading for Mathematics Placement and a Bridge Course for STEM Students," *PRIMUS*, vol. 33, no. 10, p. 1106–1120, 2023.
- [23] A. C. Graesser and H. Li, "Intelligent tutoring systems and conversational agents," in *International Encyclopedia of Education (Fourth Edition)*, Elsevier, 2023, p. 637–647.
- [24] A. Latham, "Conversational intelligent tutoring systems: The state of the art," in *Women in Computational Intelligence: Key Advances and Perspectives on Emerging Topics*, Springer, 2022, pp. 77–101.
- [25] P. K. Fink, "The role of domain knowledge in the design of an intelligent tutoring system," in *Intelligent Tutoring Systems*, Psychology Press, 2014, pp. 195–224.
- [26] A. A. Soofi and M. U. Ahmed, "A systematic review of domains, techniques, delivery modes and validation methods for intelligent tutoring systems," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 3, pp. 99–107, 2019.
- [27] H. T. Binh, N. Q. Trung and B. T. Duy, "Responsive student model in an intelligent tutoring system and its evaluation," *Education and information technologies*, vol. 26, no. 4, pp. 4969–4991, 2021.
- [28] V. Slavuj, B. Kovačić and I. Jugo, "Intelligent tutoring systems for language learning," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2015.
- [29] L. Guo, D. Wang, F. Gu, Y. Li, Y. Wang and R. Zhou, "Evolution and trends in intelligent tutoring systems research: a multidisciplinary and scientometric view," *Asia Pacific Education Review*, vol. 22, no. 3, p. 441–461, 2021.
- [30] C. Conati, O. Barral, V. Putnam and L. Rieger, "Toward personalized XAI: A case study in intelligent tutoring systems," *Artificial intelligence*, vol. 298, p. 103503, 2021.
- [31] A. A. Tawfik, J. Gatewood, J. J. Gish-Lieberman and A. J. Hampton, "Toward a definition of learning experience design," *Technology, Knowledge and Learning*, vol. 27, no. 1, pp. 309–334, 2022.
- [32] R. S. Albornoz-De Luise, M. Arevalillo-Herráez and D. Arnau, "On using conversational frameworks to support natural language interaction in intelligent tutoring systems," *IEEE Transactions on Learning Technologies*, vol. 16, no. 5, pp. 722–735, 2023.
- [33] A. Esteves, A. Filho, A. Raabe and R. Dazzi, "An Adaptive System Architecture Model for the Study of Logic and Programming with Learning Paths," in *Proceedings of the 22nd International Conference on Enterprise Information Systems*, Setúbal, Portugal, 2020.
- [34] S. Shen, Q. Liu, Z. Huang, Y. Zheng, M. Yin, M. Wang and E. Chen, "A Survey of Knowledge Tracing: Models, Variants, and Applications," *IEEE Transactions on Learning Technologies*, vol. 17, p. 1858–1879, 2024.
- [35] R. Subha, N. Gayathri, S. Sasireka, R. Sathiyabanu, B. Santhiyaa and B. Varshini, "Intelligent Tutoring Systems using Long Short-Term Memory Networks



- and Bayesian Knowledge Tracing," in *2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, Lalitpur, Nepal, 2024.
- [36] X. Bai and M. Stede, "A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring," *International Journal of Artificial Intelligence in Education*, vol. 33, no. 4, pp. 992–1030, 2023.
- [37] W. Cui, Z. Xue, J. Shen, G. Sun and J. Li, "The Item Response Theory Model for an AI-based Adaptive Learning System," in *2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET)*, Istanbul, Turkey, 2019.
- [38] E. Cosyn, H. Uzun, C. Doble and J. Matayoshi, "A practical perspective on knowledge space theory: ALEKS and its data," *Journal of Mathematical Psychology*, vol. 101, p. 102512, 2021.
- [39] R. Z. Cabada, M. L. B. Estrada and Y. H. Pérez, "Knowledge-Based System in an Affective and Intelligent Tutoring System," in *Current Trends on Knowledge-Based Systems*, Springer, 2017, p. 95–113.
- [40] R. J. C. M. Salden, V. Aleven, R. Schwonke and A. Renkl, "The expertise reversal effect and worked examples in tutored problem solving," *Instructional Science*, vol. 38, no. 3, p. 289–307, 2009.
- [41] B. Paassen, B. Mokbel and B. Hammer, "Adaptive structure metrics for automated feedback provision in intelligent tutoring systems," *Neurocomputing*, vol. 192, p. 3–13, 2016.
- [42] J. F. Pane, B. A. Griffin, D. F. McCaffrey and R. Karam, "Effectiveness of Cognitive Tutor Algebra I at Scale," *Educational Evaluation and Policy Analysis*, vol. 36, no. 2, p. 127–144, 2014.
- [43] Y. Fang, Z. Ren, X. Hu and A. C. Graesser, "A meta-analysis of the effectiveness of ALEKS on learning," *Educational Psychology*, vol. 39, no. 10, p. 1278–1292, 2018.
- [44] M. Taub, R. Azevedo, R. Rajendran, E. B. Cloude, G. Biswas and M. J. Price, "How are students' emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system?," *Learning and Instruction*, vol. 72, p. 101200, 2021.
- [45] E. Kochmar, D. D. Vu, R. Belfer, V. Gupta, I. V. Serban and J. Pineau, "Automated personalized feedback improves learning gains in an intelligent tutoring system," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020*, Ifrane, Morocco, 2020.
- [46] D. A. Dever, N. A. Sonnenfeld, M. D. Wiedbusch, S. G. Schmorow, M. J. Amon and R. Azevedo, "A complex systems approach to analyzing pedagogical agents' scaffolding of self-regulated learning within an intelligent tutoring system," *Metacognition and Learning*, vol. 18, no. 3, p. 659–691, 2023.
- [47] H. Sarrafzadeh and F. Mehdipour, "Intelligent Affect-Sensitive Tutoring Systems," in *Smart and Intelligent Systems*, CRC Press, 2021, p. 33–56.
- [48] J. Dāboliņš and J. Grundsperķis, "The Role of Feedback in Intelligent Tutoring System," *Applied Computer Systems*, vol. 14, no. 1, p. 88–93, 2013.
- [49] J. Psotka and N. S. Chen, "The new potentials for Intelligent Tutoring with learning analytics approaches," *Interactive Learning Environments*, vol. 27, no. 5–6, pp. 583–584, 2019.
- [50] K. Holstein, Z. Yu, J. Sewall, O. Popescu, B. M. McLaren and V. Aleven, "Opening Up an Intelligent Tutoring System Development Environment for Extensible Student Modeling," in *Artificial Intelligence in Education*, Springer International Publishing, 2018, p. 169–183.
- [51] S. Jiménez, R. Juárez-Ramírez, V. H. Castillo and A. Ramírez-Noriega, "Integrating affective learning into intelligent tutoring systems," *Universal Access in the Information Society*, vol. 17, no. 4, p. 679–692, 2017.



THE APPLICATION OF ARTIFICIAL INTELLIGENCE TO ASSIST PEOPLE WITH HEARING IMPAIRMENTS: AN OVERVIEW OF EXISTING SOLUTIONS

Dobrislav Drakul¹,
[0009-0001-5446-0497]

Dragan Golubović^{2*},
[0000-0003-0019-0417]

Marko Tanasković¹
[0000-0003-3592-0598]

¹Singidunum University,
Belgrade, Serbia

²Vlatacom Institute,
Belgrade, Serbia

Abstract:

Hearing impairment is a very serious problem, because it negatively affects communication between people and it also impacts their overall quality of life. Therefore, it is crucial to develop methods to help people with this problem. Although there are many hearing aids with advanced features, many analyses show that more than half of users are not satisfied. The application of Artificial Intelligence (AI) and especially Deep Learning (DL) are increasingly being mentioned as a solution. These methodologies are of recent date. Therefore, this paper provides an overview of the most important methods that use AI approaches to enable personalization and individualization of hearing aids. This primarily refers to Automatic Speech Recognition (ASR) methods, Sign Language Recognition methods (SLR), Intelligent methods for personalized hearing aids, Artificial intelligence methods in bone conduction hearing systems and Methods based on the Internet of Things (IoT). We have demonstrated the existing methods capabilities and challenges, but directions of future development are also particularly emphasized.

Keywords:

Hearing Aid Methods, Artificial Intelligence, ASR, Sign Language Recognition, IoT.

INTRODUCTION

Many studies show that almost half a billion people worldwide have some kind of hearing problem. People with hearing impairments face various problems that affect their social integration, their educational and professional opportunities and therefore it can affect their personal and work life. This is the reason why the hearing aid industry is currently in rapid progress, so millions of dollars are spent annually on hearing aid research.

Traditional hearing aids provide a certain level of support, but they have some limitations in complex acoustic environments, because a lot of people cannot understand speech in situations with background noise. Artificial Intelligence (AI) opens up new opportunities for creating personalized solutions that overcome these challenges transforming hearing aids to become smarter, more personalized, and able to adapt to different environments, offering a new level of hearing assistance that was unimaginable in previous years.

Correspondence:

Dragan Golubović

e-mail:

dragan.golubovic@vlatacom.com





AI provides innovative tools that can contribute to overcoming most challenges through technologies such as: Automatic speech recognition (ASR) and its transformation into text, speech translation into sign language using Deep Learning (DL), the development of smart hearing aids with AI functions, personalized assistive technologies to improve access to audio information, etc. With AI capabilities, hearing aids have become intelligent systems capable of analysing and adapting to different sound scenarios. AI enables not only better speech perception, but also interaction with modern technologies. This is primarily about smartphones, IoT (Internet of Things) devices and virtual assistants. This paper is related to the current state of AI hearing assistance, but it should also identify challenges and propose guidelines for further progress.

In the past, hearing aids did not have any signal processing, but their role was only to amplify the signal. But digital signal processor (DSP) expansion in the 1980s and 1990s leads to their improvement and it was the initial step toward more sophisticated hearing aids. To enhance further noise reduction in various acoustic situations, Machine Learning (ML) algorithms were introduced in the early 2000s. This was the beginning of AI in hearing aid solutions. Today, AI and DL have greatly improved real-time processing, but also the hearing aids customization.

Advanced solutions offer much more than the initial solutions and provide hearing impaired people relaxed conversations with relatives, friends or colleagues. This is enabled by the great progress of technology in recent years, especially in smaller and faster processor production, as well as in the field of multi-core and multi-processor processing. In this paper, we will focus on a few of the most important AI solutions.

A significant step is enabled by using AI in ASR technology. The main goal is to enable hearing aids to recognize, analyse and interpret speech in real time, to significantly improve sound quality, and to adapt in different environments [1] [2].

People who have very severe hearing loss still cannot communicate in the classical way. They have to use sign language to be able to communicate at all. In that case, AI is the logical choice [3] [4].

Many authors deal with intelligent methods using deep learning-based algorithms for personalized hearing aids, as presented in [5] [6] [7]. These methods selectively suppress noise while maintaining speech signals. The DL is used to provide speech intelligibility improvements. For now, at most commercially available hearing

aids, the beamforming (spatial selectivity) is often used to suppress a noise.

It is important to notice that AI also enables Bone Conduction technology to become more precise, adaptive and intelligent for different users enabling them have an easier life resulting in advanced possibilities of their interacting with the environment [8] [9]. The advancement of this technology is of particular importance.

AI methods with the IoT enable the significant improvement. We can connect different devices and sensors in a network and we can analyse and adjust settings in real time, enabling better recognition of sound signals, improved security and greater autonomy [10] [11] [12].

The paper is divided into several parts. The main principles of hearing assistance are presented in Section 2. Section 3 is the main part, where we analyse the most important existing methodologies to assist people with hearing impairments. In Section 4, the effectiveness of the existing methods and the direction of future development are shown. Section 5 is related to conclusions.

2. HEARING ASSIST FUNDAMENTALS

There are several levels of hearing loss. We can say that the hearing threshold of 20 dB is actually the limit that indicates that a person has hearing problems. These problems can be mild or moderate, but they can also be very serious (severe or profound). It is also important to emphasize that the problems can exist in both ears, but they can also be asymmetrical, where they exist in only one ear. All these problems are analysed and solved depending on the specific situation, but in recent years they have begun to be methodologically solved at a significantly higher level. Figure 1 shows typical hearing thresholds.

Even mild hearing loss people need help and some kind of assistance is necessary. They can still use spoken language, but aids, cochlear implants, and other assistive devices are required. If hearing is still profoundly impaired, people often use sign language for communication. Hearing aids can reduce environmental and enhance speech by using intelligent AI algorithms. Hearing problems can be classified into several categories.

- Difficulty understanding speech
- Misunderstanding information
- Reduced ability to participate in conversations

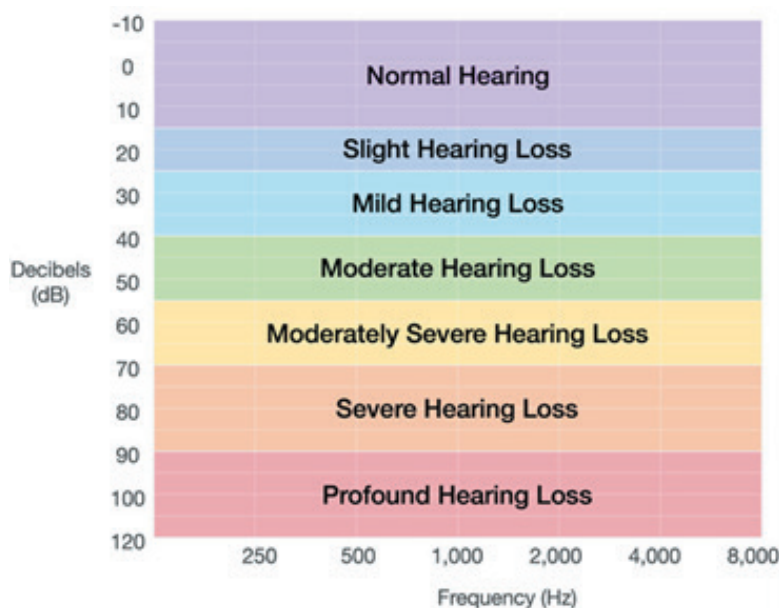


Figure 1. Hearing threshold typical values [13]

Assistive listening devices, which include hearing aids and hearing aid accessories are used, so that people can fully engage in communication, even if their hearing is severely impaired. The sound is amplified and, at the same time, the noise is suppressed, so the communication is much easier.

AI and ML are used in the Internet of Things (IoT) applications, and they are also applicable in hearing aid purposes. This paper is related to the systematic review of existing AI methods in hearing aids. We present the most significant challenges, as well as the things that pose a challenge in the future. AI model can learn user preferences and behaviours. If the user frequently changes settings in certain situations, the device can automatically apply those changes and can adapt to it, by analysing how the user reacts to different sounds.

3. THE METHODOLOGIES TO ASSIST PEOPLE WITH HEARING IMPAIRMENTS

AI hearing aids are based on complex technology, intelligent algorithms and sophisticated signal processing. It is necessary to highlight three core components of such systems to provide a more personalized and hassle-free hearing experience, as presented in Figure 2:

- **Artificial intelligence (AI)** is the most important part because it represents computer technology in order to make some devices do human intelligence things. Here, for example, it is image recognition or language understanding.

- **Machine learning (ML)** enables various algorithms to improve themselves from the user experience. These algorithms process large amounts of information and find patterns that help them make predictions or decisions.
- **A Deep Neural Network (DNN)** tries to work like the human brain. It is a kind of machine learning algorithm consisting of many layers that work together to solve complex problems, like recognizing faces or translating languages.

The merging of AI technologies and hearing aids represents an advanced solution to solve many hearing problems for the millions that live with hearing impairment around the world. The main goal is to enhance real-time listening using AI engines, helping people to personalize their sound experience. In this section, several of the most important existing solutions are described. Figure 2 shows AI and ML description model.



Figure 2. AI and ML description model

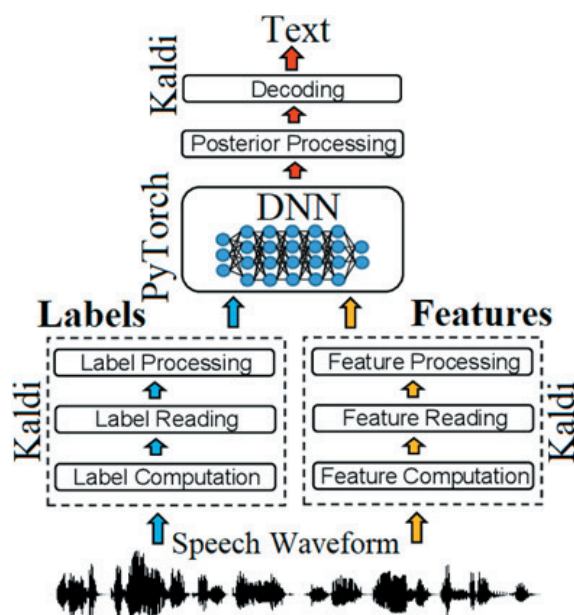


Figure 3. Automatic Speech Recognition (ASR) description model

3.1. AUTOMATIC SPEECH RECOGNITION (ASR) METHODS

ASR is a methodology based on AI and ML, according to which human speech is converted into text by using different algorithms for its recognizing and processing. ASR depends on various components such as preprocessing, speech classification, various feature "extraction" techniques, classification, databases, and system performance. From the technology perspective, speech recognition has benefited from advances in DL and big data. The advances are evidenced not only by a lot of published scientific papers, but more importantly by the worldwide industry adoption of a variety of deep learning methods. In recent years, there has been rapid development, with popular applications for real-time captions (TikTok and Instagram), for podcast transcriptions (Spotify), for meeting transcriptions (Zoom), YouTube and more. Figure 3 shows typical ASR description model.

Also, one of the basic components of AI speech processing is a technique known as Natural Language Processing (NLP). People and the AI system are in the most natural human form interaction. In that way, it enables to extract language characteristics based on a sound recording. The synthesis and analysis of human speech is performed by training neural networks, i.e. DNN with a large amount of data.

The YouTube application uses automatic speech recognition to generate translations of videos. We can turn on Subtitle/CC option in the settings and video clip is processed in real-time in order to have a translation. The ASR enables people with partial hearing loss not only to communicate digitally, but also to communicate in real life, because with the advent of portable devices, these possibilities are always available to them.

There are plenty of toolkits that allow for easy speech recondition using different programming languages. Some of the most important open-source tools can be seen in Table 1.



Table 1. The most important open-source speech recognition tools

Tool	Programming languages	Trained models
CMU Sphinx	C, Python, Java, others	English and 10 other languages
Kaldi	C++, Python	English
Julius	C, Python	Japanese
ISIP	C++	Digits only

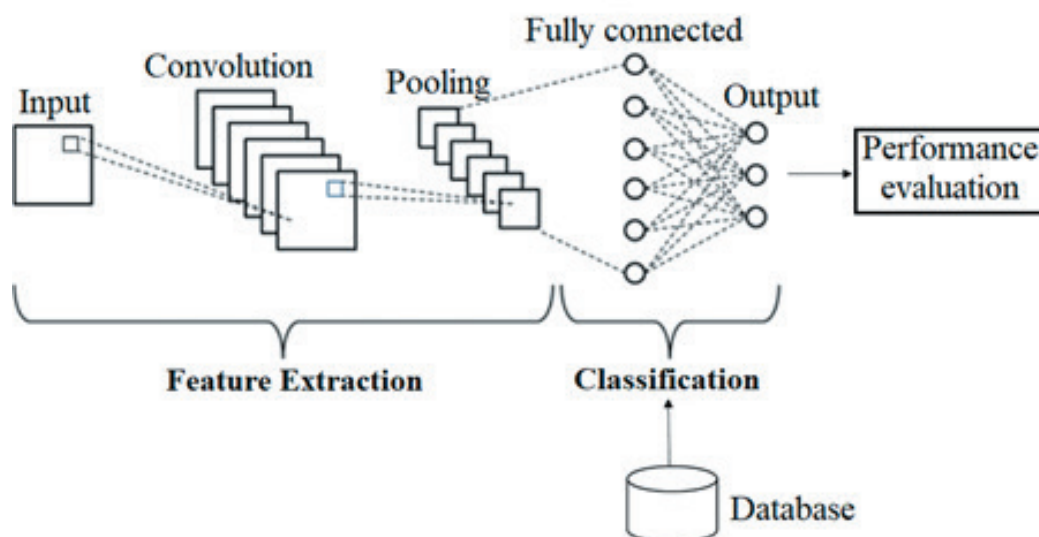


Figure 4. Sign language recognition model using AI

3.2. SIGN LANGUAGE RECOGNITION METHODS

People who have severe hearing problems can use sign language as a way of communication. Recently, the field of computer vision and DL has been developing a lot. As a result, vision-based programs are being developed that also enable the translation of sign language to text. It is necessary to take video sequences and extract temporal and spatial features from them. Then Inception process starts using CNN (Convolutional Neural Network) for recognizing spatial features. Finally, RNN (Recurrent Neural Network) is used to train on temporal features. The dataset used is Sign Language Dataset that corresponds to some speak language. Figure 4 shows a sign language recognition model using AI.

Mobile and desktop applications use AI to translate speech into text format. But real-time translation to sign language is also possible. Sign language translation systems are based on computer vision and DNN, so that hand gestures and facial expressions can be recognized. The combination of different sensor inputs can significantly increase the accuracy and efficiency of the system.

The speech translation into sign language consists of:

- Speech recognition where speech is recognized using the ASR
- Motion analysis where cameras and sensors track hand gestures and facial expressions.
- Translation into sign language where DNN map recognized speech elements to sign language.

Sign language translation applications can be integrated with smartphones, computers and devices for easier communication. Therefore, sign language people can communicate with people who speak verbally through mobile applications that enable mutual communication in real time, often by using Cloud-based services.

3.3. INTELLIGENT METHODS FOR PERSONALIZED HEARING AIDS

Even when using hearing aids, people may have understanding problems in public places, such as stadiums, cafes, etc. In that case, it is also necessary to use some



advanced algorithms to suppress unwanted environmental noise. AI can analyse sound and automatically adjust the hearing aid to focus on important sounds, improving speech recognition and the personalization of the experience is also enabled. AI enables hearing aids to become a much smarter, more adaptive, improving user’s ability to communicate and enjoy everyday activities. Table 2 shows key components of AI hearing aids, while Figure 5 shows the display of a typical AI hearing aid and one example of a hearing aid application.

3.4. AI IN BONE CONDUCTION HEARING SYSTEMS

Bone Conduction (BC) technology thar uses vibrations to transmit sound directly through the bones of the skull to the inner ear, bypassing the middle and outer ear.

Sound is converted into vibrations that are transmit- ted by special BC devices through the temporal bone to the cochlea (inner ear). This technology enables listening even for people with middle ear damage or a complete lack of the external ear. AI in these devices enables the adaptation of sound vibrations according to the user's frequency needs. For example, AI algorithms can adjust the intensity of vibrations to optimize sound reception without causing discomfort. AI in BC technology is used to improve sound quality, filter noise and adapt to user needs. Intelligent noise filtering and speech enhance- ment are enabled in that case and AI algorithms analyse the vibrations transmitted through the skull and dis- tinguish useful sounds (speech) from unwanted noise. Figure 6 shows an example of the BC hearing system.

Table 2. Key hardware and software components of AI hearing aids

Component type	Component role
Microphones	Input for capture sound from the environment and to form o audio signal
Processors	Use different AI algorithms to analyse and process input audio signal
Receivers	They are used to send processed audio signal to user’s ear
Connectivity module	It is used to communicate with other devices, such as tablets, smartphones, smartwatches, etc.
Software	It is used to execute AI algorithms and to start user interfaces, often via different mobile applications



Figure 5. (a) AI hearing aids (b) An example of hearing aid Application



Figure 6. An example of the BC hearing system



3.5. METHODS BASED ON THE INTEGRATION WITH THE INTERNET OF THINGS

Interaction with other devices and technologies is possible by connecting with smartphones, computers or other devices. AI enables hearing aids to recognize and integrate different forms of sound, such as phone calls, music, sound signals from the TV, and to process and adapt them accordingly. Also, they can be synchronized with speech recognition applications improving the user experience in everyday interactions. One of the advantages of AI and IoT integration is better safety because IoT devices can recognize warning and sound signals (alarms, sirens) and notify the user in a visual or vibrational way. Figure 7 shows the IoT hearing aid system.

4. THE EFFECTIVENESS OF THE EXISTING METHODS AND THE DIRECTION OF FUTURE DEVELOPMENT

Personalisation and device setting adjustment is of crucial importance. Personalization is influenced by multiple factors, and it is important that the manipulation is not too complex, regardless of the user's level of education or skill. Psychosocial and cultural factors also play an important role. A key metric used in many studies is the acceptance rate of the proposed solutions. The overall acceptance rate for all proposed solutions was more than 80% in the last 20 years. This growing acceptance shows that the AI-enabled hearing aid applications are successfully accepted solutions and have progressively improved performance over time.

So, the conclusion is that the future of AI and hearing aids is bright. Context-aware personalization is another future direction, with advanced solutions that automatically identify a user's operating audio environment and adjust settings accordingly. Effective feedback mechanisms, such as surveys and continuous feedback loops, will be useful for refining personalization solutions based on a user's experiences. Integration with portable devices is also one of the most important tasks in order to ensure further personalization enhancement. The ultimate goal of AI hearing aid technology is to improve audibility, speech perception, and comfort in listening.

5. CONCLUSION

The latest generations of hearing aids will focus on lowering costs to make devices more accessible to more users around the world. The revolution with AI capabilities in real-world scenarios, unlike traditional methods, opens the way for personalized experiences in noisy and dynamic environments. In this paper, we have shown that there are several hearing aid solutions that provide high-quality audio to the users, making them comfortable. It is evident that AI-enabled hearing aid applications have begun more personalized, efficient, and satisfying for users. Also, by using the advantages of wireless communication, hearing aids have become almost invisible and look like a piece of jewellery. The proposed solutions are user-oriented because only newer and more advanced solutions provide a satisfactory experience to users, because their living standards and expectations have also improved in recent years.



Figure 7. IoT and smart AI hearing aid interaction



REFERENCES

- [1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, Feb. 1989.
- [2] Y. Shi, M.-Y. Hwang and X. Lei, "End-to-end Speech Recognition Using a High Rank LSTM-CTC Based Model," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.
- [3] K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision," in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018.
- [4] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling and S. Narayanan, "Real-time sign language detection using human pose estimation," in *Computer Vision-ECCV 2020 Workshops*, Glasgow, UK, August 23–28, 2020.
- [5] P. U. Diehl, Y. Singer, H. Zilly, U. Schönfeld, P. Meyer-Rachner, M. Berry, H. Sprekeler, E. Spengel, A. Pudzuhn and V. M. Hofmann, "Restoring speech intelligibility for hearing aid users with deep learning," *Sci Rep* 13, p. 2719, 2023.
- [6] A. H. Andersen, S. Santurette, M. S. Pedersen, E. Alickovic, L. Fiedler, J. Jensen and T. Behrens, "Creating Clarity in Noisy Environments by Using Deep Learning in Hearing Aids," in *Seminars in hearing*, 2021.
- [7] N. Z. Tasnim, A. Ni, E. Lobarinas and N. Kehtarnavaz, "A Review of Machine Learning Approaches for the Personalization of Amplification in Hearing Aids," *Sensors*, vol. 24, no. 5, p. 1546, 2024.
- [8] S. E. Ellsperman, E. M. Nairn and E. Z. Stucken, "Review of Bone Conduction Hearing Devices," *Audiology research*, vol. 11, no. 2, p. 207–219, 2021.
- [9] Y. Gupta, T. Choudhury, P. Kumar and S. Kumar, "Bone Conduction Auxiliary and Tactics for Man Machine Interface for Hearing Impaired Users," in *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, Paris, France, 2018.
- [10] İ. Kök, F. Y. Okay, Ö. Muyanli and S. Özdemir, "Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey," *IEEE Internet of Things Journal*, vol. 10, no. 16, pp. 14764–14779, 2023.
- [11] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali and M. Guizani, "A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1646–1685, 2020.
- [12] X. Ma, T. Yao, M. Hu, Y. Dong, W. Liu, F. Wang and J. Liu, "A Survey on Deep Learning Empowered IoT Applications," *IEEE Access*, vol. 7, pp. 181721–181732, 2019.
- [13] K. Taylor and W. Sheikh, "Automated hearing impairment diagnosis using machine-learning: An open-source software development undergraduate research project," *Comput. Appl. Eng. Educ.*, vol. 32, p. 22724, 2024.



DESIGN AND DEVELOPMENT OF DIGITAL TWINS IN SIMULINK

Dragan Kreculj¹,
[0000-0003-3268-4024]

Đorđe Dihovični¹,
[0000-0003-0961-2540]

Nada Ratković Kovačević^{1*},
[0000-0001-6398-4391]

Aleksandra Mitrović^{1,2},
[0000-0002-5183-6276]

Maja Gaborov³
[0000-0002-3810-6156]

¹The Academy of Applied Studies
Polytechnic,
Belgrade, Serbia

²University "Union - Nikola Tesla",
Faculty of Information Technology and
Engineering,
Belgrade, Serbia

³University of Novi Sad,
Technical Faculty "Mihajlo Pupin",
Novi Sad, Serbia

Correspondence:

Nada Ratković Kovačević

e-mail:

nratkovickovacevic@politehnika.edu.rs

Abstract:

A digital twin is a real-time, dynamic digital representation of a physical system, enabling continuous monitoring, simulation, and analysis. It plays a crucial role in optimizing system performance, predictive maintenance, and decision-making across various engineering domains. Simulink, a MATLAB-based software environment, is widely utilized for modelling and simulating complex dynamical systems through block diagrams.

This paper aims to explore the development process of digital twins within Simulink, highlighting its capabilities for system-level design and validation. Simulink provides engineers and researchers with powerful tools to create, simulate, and optimize digital twins of physical systems. By integrating real-time data, these models enable accurate performance evaluation, fault detection, and predictive analytics, making them essential for industries such as aerospace, automotive, robotics, civil engineering, and manufacturing.

The study outlines key methodologies for developing digital twins in Simulink, emphasizing their role in enhancing system reliability, improving operational efficiency, and accelerating innovation. Additionally, it discusses Simulink's features, such as its multi-domain modelling capabilities, simulation accuracy in real time, and integration with real-world data sources, which contribute to the improvement of digital twin technology. It is demonstrated that Simulink-based digital twins facilitate engineering decision-making, streamline system optimization, and drive technological advancements in industrial applications.

Keywords:

Digital Twins, Design and Development, Modelling and Simulation, Optimization, Simulink.

INTRODUCTION

A digital twin is a virtual representation of an entity (e.g. object, process, system, machine, or device) designed to reflect physical properties and behaviours accurately. It is updated from real-time data and uses simulation, machine learning (ML) and automated reasoning to help make decisions. A digital twin is a virtual replica that allows for real-time monitoring, simulation and analysis. The digital twin enables performance optimization, better decision-making, faster problem solving and more efficient management of resources. The properties of the digital twin are crucial for its successful use in various industries, such as aerospace, automotive, civil engineering, maintenance, education, manufacturing, robotics, for system-level design. [1]



A digital twin is a virtual replica of a product, process, or system—whether in operation or under development. For assets in operation, the digital twin reflects the current condition and incorporates pertinent historical data. Digital twins are used for evaluating an asset's current state and to predict future behaviour, refine control systems, and optimize the operation of systems, either objects or processes. In the development procedure, the digital twin acts as a model of a product to be built, process, or system that facilitates development, testing and validation. [2]

MATLAB [3] and its Simulink [4] toolbox (both products of MathWorks, USA) offer platforms for the creation, simulation, verification and ultimately implementation of specific digital twins. Simulink offers a set of tools and features to create, simulate, and optimize digital twins of real systems. The combined physics-based modelling, advanced data analytics and Artificial Intelligence (AI) tools, and options for easy deployment (Programmable Logic Controllers - PLCs, embedded systems, web, cloud), enable engineers and designers to obtain directed design of digital twins that enhance the understanding, operation and maintenance of complex entities. [2]

Digital twin platforms are changing various commerce sectors with powerful capabilities that facilitate data-driven choices, predictive maintenance, and real-time monitoring. These platforms provide unique perspectives that drive efficiency, creativity, and environmental awareness in industries, including manufacturing, healthcare, energy, transportation, construction, and the economy in general. [5]

Digital Twins can also be perceived as models of real or virtual entities (objects, components, systems or processes) which extend options for operation simulation and evaluation of different scenarios.

2. MODELLING AND SIMULATION

Modelling and simulation are well-established and often exploited in engineering for design and automation of processes [6]. System or model identification and automatic control are valuable in contemporary engineering and industrial automation. Simulink [4] is a specialized software that enables efficient modelling of the design, operation simulation and analysis of complex systems. In the following part, the key aspects of how to use Simulink for automated modelling are discussed briefly. Also, the key components of MATLAB and Simulink which are exploited in modelling and simulation of automated systems are concisely explored.

Simulink is a software environment for modelling, simulation and analysis of a multi-domain dynamical systems described by block diagrams. Simulink uses a graphical description (by Simulink Blocks) to visually represent system components and their interactions. For example, modelling a production conveyor belt uses blocks for its conveyor, robotic arms and controls to compose a description and model the entire system. Simulink is user friendly since block diagrams of systems can be made by drag-and-drop of standard modules and components selected from a menu.

For control logics, the state flow diagrams are used to define and model complex decision-making processes and behaviour, depending on the overall state in automated systems. This determines the control logic in the existing system. For example, in an automated traffic light system, timely state transitions and vehicle detection in real time can be designed, clearly defining the engineering logic for each state (or colour) of the light.

Specific tools and features in Simulink help engineers to create, implement and exploit Digital Twins. The development of a digital twin in Simulink has the ability to significantly improve the simulated processes and analysis of physical systems. The benefits of using digital twins in Simulink are the following: improved efficiency, reduced costs and enhanced innovation. The key features of Digital Twin models in Simulink are expressed as real-time data visualization and predictive maintenance. Both digital twins and Simulink are applied in various industries.

Simulink offers an appropriate set of tools and features to create, simulate, optimize Digital Twins of real systems. Simulink provides several utilities which are aimed to create precise and efficient digital twin models that can be used for various applications, such as predictive maintenance, performance optimization and simulation of different scenarios. These tools and toolboxes allow for real-time simulation, predictive maintenance, optimization and visualization of system behaviour. Options provided by Simulink are:

1. Simulink 3D Animation: Visualize and simulate digital twin in 3D, providing an intuitive representation of the system's behaviour;
2. Simulink Real-Time: Simulink Real-Time is used to run real-time models for hardware-in-the-loop (HIL) testing. This module allows one to run the model on hardware in real-time and interface it with real-world data. It is important for hardware-in-the-loop (HIL) simulation, where the test model (or system emulation) is implemented



in a real-world environment. This is essential for automating processes in aerospace or automotive applications, such as assessing the accuracy of control algorithms in real time while interacting with real, existing, operating hardware;

3. Stateflow [7]: Stateflow is a product of Math-Works that provides a graphical language which includes state transition diagrams, flow charts, state transition tables, and truth tables. Stateflow fosters modelling of complex logic and behaviour in the digital twin, such as state machines, events and transitions occurring in the physical system;
4. MATLAB Integration: MATLAB is used in parallel to Simulink for data analysis, visualization and more sophisticated algorithms that may not be directly implemented within Simulink. Integration with MATLAB provides additional sensor data analysis and system behaviour optimization;
5. Simscape [8]: The Simscape enables developers and designers to rapidly create models of physical systems within the Simulink environment. This module extends Simulink to simulate physical systems, including mechanical, hydraulic, and electrical components, or thermal systems. For example, modelling a robotic arm in the Simscape, which requires precise control of joints and movements of segments provides a simulation of relevant physical interactions, making it easier to optimize overall performance. Simscape module is suited for simulating the behaviour of components and entities that formed a digital twin.

The MATLAB modules that complement the use of Simulink in creating digital twins are Statistics and Machine Learning Toolbox and Deep Learning Toolbox.

Additional resources provided by MATLAB and Simulink modules that can help in the design and development of a digital twin are Signal Processing Toolbox and Control System Toolbox.

Simulink's Control System Toolbox offers predefined PID controllers (having Proportional, Integral and/ or Derivative action - PID) for automating specific processes. PIDs have editable parameters providing ease-of-use in implementation.

Simulink's Signal processing tool is used for filtering and processing signals in automated systems. In an automated quality control system, data acquisition and processing can be modelled to detect defects in products using available image processing techniques.

All these modules and toolboxes create a powerful set of tools for developing digital twins that are to be used in various industries and applications e.g. to predict the performance of a drilling rig in real time, [9].

Automation of management strategies is realized by building adaptable management systems, which responds and adapts to data in real time. For example, in a smart HVAC (Heating, Ventilation and Air-conditioning) system, control logic can be adjusted based on room occupancy and weather conditions to optimize energy consumption.

Simulink software allows running simulations under different conditions to evaluate the performance of selected systems. In an automated warehouse, different scenarios of inventory levels and order fulfilment rates can be simulated, in order to evaluate the performance of the logistics management system.

MATLAB scripts can be implemented to automate tasks in Simulink. For example, scripts can be created that automatically adjust model parameters based on optimization results, governing/ driving/ directing the model tuning process for an automated process control system. Simulink options are used to generate automated reports from simulations, summarizing performance metrics, model parameters, and results. This can be particularly useful in projects where compliance of documentation (with standards), or analysis, is required.

3. DESIGN OF DIGITAL TWINS

The development and design of a digital twin represent a dynamic and innovative approach to the simulation and optimization of physical objects and systems. A digital twin replicates a physical entity or system, enabling real-time data analysis. [10]

Despite the variety of digital twin use cases, there are similar strategies to make successful implementation of digital twins. These strategies are usually the following: define clear objectives and scope, design and build, perform testing and validate, deploy and operate, monitor and update. [2]

These strategies are sometimes called the digital twin product development stages, e.g. in [5] or the key steps in the development of a digital twin as in [11]: defining goals, collecting data, creating models, integration with IoT (Internet of Things), analysis and simulation, optimization and improvement.



The first step in the development of a digital twin is to clearly define the goals and purpose of a digital twin which is being built. Then it is necessary to collect all relevant data from various sources, including sensors, IoT devices and historical data [12].

Based on the collected data, a virtual model is then created. Specialized software such as CAD (Computer-Aided Design) and simulation tools are used to create 3D models. For example, Autodesk Fusion 360 can help develop a model – a 3D visualization of a physical system and ultimately design and build its digital twin or the real object.

The digital twin connects to IoT platforms in order to receive data in real time. For example, for connecting physical objects with digital models platforms such as Microsoft Azure IoT, or AWS IoT can be used.

After integration, analyses can be carried out based on simulations. These can include predictive maintenance, when data is analysed to predict failures before they occur. Based on detailed analyses, the digital twin can help significantly improve, enhance, and optimize the overall performance of the system under consideration.

Validation and verification are critical steps in ensuring the digital twin potential to accurately reflect a physical counterpart and its expected performance. By using the high-integrity verification workflow, engineers use simulation-based testing and static analysis to detect some defects, damages and to shorten the time to the final market. Moreover, the priority is to adhere to standards and maintain high product quality.

The stages of digital twin product development are presented in Figure 1. [5]

The more detailed steps to design effective digital twin solution for manufacturing are given in Figure 2. [11]

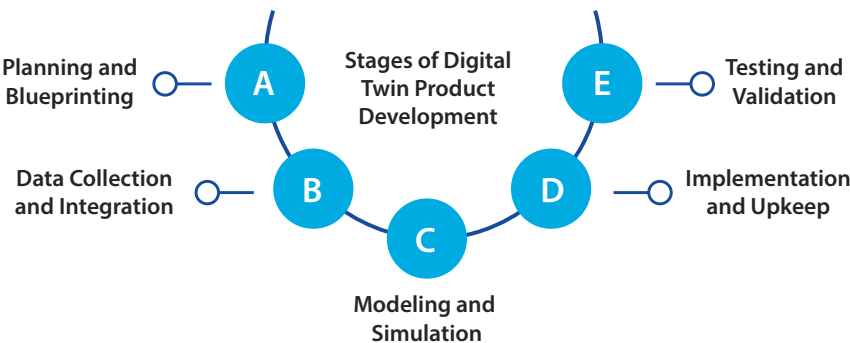


Figure 1. Digital Twin Product Development Stages or Phases¹

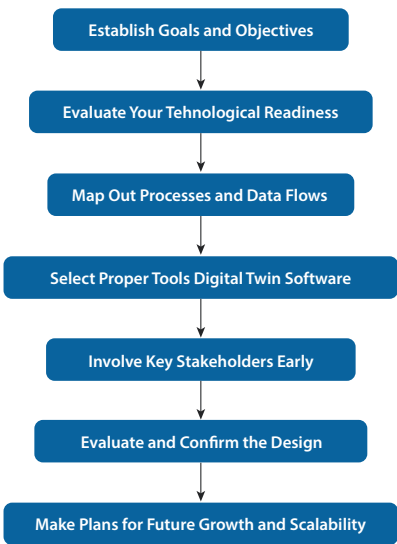


Figure 2. Steps to Design an Effective Digital Twins Solution for Manufacturing²

1 Source: <https://a.storyblok.com/f/122804/1600x1000/2c4914f8fa/stages-of-digital-twin-product-development.webp>
2 Source: <https://a.storyblok.com/f/122804/3201x3868/a3467ad142/steps-to-design-effective-digital-twin-solutions-for-manufacturing.webp>



Development of a digital twin is a process which usually lasts several weeks to months, depending on its complexity and use intended [5].

Digital twins for Cyber-Physical Systems (CPSs) have been built using different methods, e.g. application of model-based systems engineering approach and ML techniques. The CPS model refers to a virtual replica or representation of the CPS, which simulates the CPS behaviour. [13]

The digital twin of CPS consists of two main components:

1. Digital twin model; and
2. Digital twin capability.

The elements of digital twin and their interactions are displayed in Figure 3. [13]

Digital twins created in Simulink can be exploited to enhance the design, monitoring, and maintenance of complex technological systems. By leveraging real-time data and advanced simulation capabilities, engineers can acquire deeper insights that drive operational efficiency, enhancements and innovation in many areas.

4. THE PERFORMANCES OF DIGITAL TWINS

The performance of the digital twin is crucial for its successful exploitation in various use cases and industries. Here are briefly described some aspects of the digital twins' implementation and optimisation achieved by these.

- I. Real-time Data Processing: The rate and efficiency of processing the data collected from the sensors significantly affect the performance of the digital twin. The latency of the system will be reduced using edge computing technologies since data processing is then performed closer to the data source;
- II. Data Quality: The accuracy and relevance of the data used by the digital twin is important for its performance. Implementing a system for collecting, testing, validating and cleaning data can help to ensure its high quality;

III. Model Complexity: More complex models provide more detailed analysis, albeit digital twins' operation could be made slower. Developing simplified models for quicker analysis and more complex models for more accurate and detailed analysis can improve overall performance;

IV. Interoperability: The capability of the digital twin created to communicate with different systems and platforms can affect its effectiveness. Using standardized protocols and APIs (Application Programming Interfaces) makes it possible to improve interoperability;

V. Simulation and Prediction: The accuracy of simulations and predictions directly affects the implementation and applications of the digital twin. Applying ML models that are trained on historical data can improve predictive capabilities and prediction accuracy;

VI. Data visualization: Data collected can be used to support relevant automated decision-making. The interactive graphs and visualizations have the ability to improve the information extraction from data;

VII. Scalability: The digital twin is adapted to scale (or increase) in size and time with respect to specified demands. Using cloud computing resources enables easy scaling in accordance with the requirements and needs.

By connecting digital twins of the machines or assets during operation with relevant data, algorithms and models, it is possible to make better decisions, improve processes and reduce the overall losses and total cost. Some research papers indicate that investment in custom digital twin development can reduce operating costs by up to 30% and boost productivity by 20% [5].

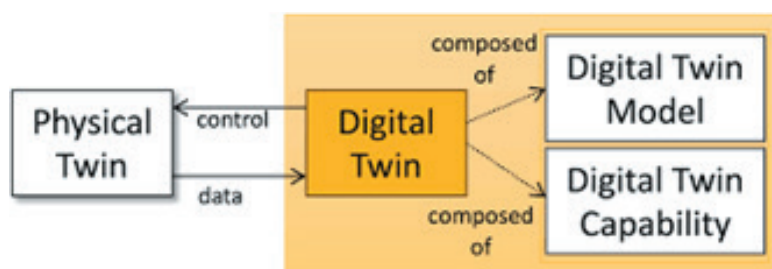


Figure 3. Digital Twin for Cyber-Physical Systems³

³ Source: <https://arxiv.org/html/2407.18779v1/extracted/5757201/images/OverviewDT.png>



5. CONCLUSION

A digital twin operates by digitally replicating a physical asset in a virtual environment, accurately mirroring its functionality, features, and behaviour. Digital twins represent a powerful tool for advancing innovation and optimizing performance across various industries. By linking the physical and digital realms, these models provide enhanced insight into system performance, enable the anticipation of potential issues, and support informed decision-making. Moreover, the integration of artificial intelligence, machine learning, and data analysis facilitates the creation of specific digital simulation models, offering a novel approach to product lifecycle management.

Simulink is a versatile and comprehensive platform that comprises specialized toolboxes and modules for modelling, simulation, and analysis of dynamic systems. It is extensively used in the development of digital twins, allowing for the creation and testing of models with high precision. By employing various simulation blocks and incorporating real-world or real-time data, robust models can be established to optimize performance and predict potential problems. This methodology is widely applied in predictive maintenance, fault detection, and overall system optimization in industries such as manufacturing, automotive, aerospace, and energy.

The development of digital twins in Simulink encompasses modelling, data integration, simulation, analysis, validation, and optimization. This process provides deeper insights into the operation of physical systems and facilitates continuous performance improvement. Several tools exist in Simulink that enable designers or developers to model complex systems specific for selected industries, to test them virtually, predict their behaviour and optimize operations, while exploiting real-time data for permanent system improvement and minimization of cost.

The objective of this work was to present a concise overview of the Simulink modules and tools used for digital twin development and implementation. Ultimately, this approach contributes to enhanced operational efficiency, cost reduction, and system resilience, driving continuous technological advancement in engineering applications.

REFERENCES

- [1] D. Kreculj, Đ. Dihovični, N. Ratković Kovačević, G. Nestorović, and P. Jakovljević, "Digital Twins in Aviation," presented at the Int. Symp. on Sustain. Aviation – ISSA 2024, Belgrade, Republic of Serbia, October 30 - November 01, 2024, unpublished, in preparation for publication
- [2] "What Is a Digital Twin?." MathWorks.com. Accessed: Mar. 08, 2025. [Online.] Available: <https://www.mathworks.com/discovery/digital-twin.html>
- [3] "MATLAB." MathWorks.com. Accessed: Mar. 30, 2025. [Online.] Available: <https://www.mathworks.com/products/>
- [4] "Simulink." MathWorks.com. Accessed: Mar. 30, 2025. [Online.] Available: <https://www.mathworks.com/products/simulink.html>
- [5] VP Ankitha. "Digital Twin Development Time." toobler.com. Accessed: Feb. 15, 2025. [Online.] Available: <https://www.toobler.com/blog/digital-twin-development-time>
- [6] C. Petrescu and V. David, Eds. *Modeling and Simulation in Engineering*. Basel, Switzerland: MDPI, 2022. Accessed: Mar. 30, 2025. DOI: 10.3390 [Online.] Available: https://mdpi-res.com/bookfiles/book/6451/Modeling_and_Simulation_in_Engineering.pdf
- [7] "Stateflow." MathWorks.com. Accessed: Mar. 15, 2025. [Online.] Available: <https://www.mathworks.com/products/stateflow.html>
- [8] "Simscape." MathWorks.com. Accessed: Mar. 18, 2025. [Online.] Available: <https://www.mathworks.com/products/simscape.html>
- [9] W. Johnson from Renoir Consulting LLC. "Using a Digital Twin to Predict the Real-Time Performance of a Drilling Rig." MathWorks.com. Accessed: Mar. 18, 2025. [Online.] Available: <https://www.mathworks.com/company/technical-articles/using-a-digital-twin-to-predict-the-real-time-performance-of-a-drilling-rig.html>
- [10] D. Kreculj, A. Mitrović, Đ. Dihovični, N. Ratković Kovačević, and A. Vučićević. (Feb. 2025). Microsoft Azure DigitalTwins. In *Proc. of the 1st Int. Conf. "Annual conference on Challenges of Contemporary Higher Education" (ACCHE)*, P. Tanović Ed., Kopaonik, Republic of Serbia. pp. 97-102. Accessed: Mar. 08, 2025. [Online.] Available: https://acche.rs/ACCHE_2025/radovi/electrical/13.pdf
- [11] VP Ankitha. "How to Design, Develop, and Deploy Effective Digital Twin Solutions for Manufacturing Industry." toobler.com. Accessed: Feb. 15, 2025. [Online.] Available: <https://www.toobler.com/blog/digital-twin-solutions-manufacturing>
- [12] F. Akbar. "How Digital Twins and IoT Work Together [With Example]." toobler.com. Accessed: Mar. 08, 2025. [Online.] Available: <https://www.toobler.com/blog/digital-twin-iot>
- [13] S. Ali, P. Arcaini, and A. Arrieta, "Foundation Models for the Digital Twin Creation of Cyber-Physical Systems," 2024, arXiv:2407.18779v1 [cs.SE]. Accessed: Mar. 15, 2025. [Online.] Available: <https://arxiv.org/html/2407.18779v1>



VALIDATION OF THE FDM-BASED ADDITIVE MANUFACTURING METHOD FOR RAPID PROTOTYPING USING THE EXAMPLE OF THE EIFFEL TOWER MODEL

Dragan Cvetković^{1*},
[0000-0002-4405-3775]

Nikola Vorkapić²,
[0000-0001-5081-4828]

Ivana Cvetković²,
[0000-0002-2773-5643]

Srđan Trajković³,
[0009-0005-1712-1065]

Aleksandar Cvetković⁴
[0009-0003-3300-9845]

¹Singidunum University,
Belgrade, Serbia

²Faculty of Mechanical Engineering,
Belgrade, Serbia

³Academy of Applied Science Politehnika,
Belgrade, Serbia

⁴Jaroslav Černy Institute for
Water Management,
Belgrade, Serbia

Correspondence:
Dragan Cvetković

e-mail:
dcvetkovic@singidunum.ac.rs

Abstract:

Modern rapid prototyping techniques, enabled by computer-aided manufacturing, allow for the efficient and accurate materialization of highly complex 3D objects designed within CAD software environments. These methods significantly reduce manufacturing time and eliminate the need for extensive post-processing. Following the validation of a digital prototype within a CAD system, the physical model is manufactured using appropriate layer-by-layer rapid prototyping technologies, including additive, subtractive, formative, or hybrid manufacturing methods.

This paper presented the use of commercial CAD tools for the modeling of geometrically complex 3D objects and their preparation for manufacturing via the Fused Deposition Modeling (FDM) additive manufacturing process. The Eiffel Tower model was selected as a representative example due to its complex structure and large dimensions. To facilitate the fabrication process, the model was divided into simpler, yet geometrically complex segments. The results demonstrate that the integration of CAD-based design with rapid prototyping techniques enables the production of large, detailed components with high dimensional accuracy and satisfactory surface quality.

Keywords:

Rapid Prototyping, Additive Manufacturing, Prototype, 3D Printing.

INTRODUCTION

In earlier times, product designers and engineers created improvised models using basic tools to demonstrate the conceptual design of a product. The manufacturing of functional prototypes required the same processes as those used for final products. Traditional manufacturing methods, such as plastic injection molding, demanded the production of expensive tooling, making small quantities of functional prototypes prohibitively costly. Rapid Prototyping emerged in the late 1980s with the introduction of the first machine by 3D Systems, which operated based on the SLA (Stereolithography) method [1–3]. Shortly thereafter, the first machines using the principle of fused deposition of material (Fused Deposition Modeling – FDM) were introduced. Since the deposited material can take the form of photopolymer, powder, or filament, more than ten different additive manufacturing methods have since been developed [4, 5].

Modern rapid prototyping methods rely on computer-aided manufacturing, enabling fast materialization of complex 3D objects created in various CAD software packages, without the need for additional post-processing. Their accessibility lies in their ease of use and the ability to work with a wide range of thermoplastics, making them suitable not only for industrial applications, but also for use in education, architecture, and the arts. Most additive manufacturing system manufacturers offer dedicated software with user-friendly instructions, allowing for simple and efficient operation. This paper presents the process of manufacturing a scale model of the Eiffel Tower, an iconic architectural structure, using FDM technology. Due to the large dimensions of the model relative to the limited build volume of the 3D printer, a segmentation approach was employed—dividing the model into multiple smaller parts of complex geometry, which were manufactured individually and later assembled into a complete structure. Special attention was given to optimizing the slicing process, selecting the appropriate orientation, and adjusting printing parameters to achieve the desired strength, dimensional accuracy, and aesthetic quality of the final construction.

2. MODEL DESIGN FOR VERIFYING THE FDM MANUFACTURING METHOD

This chapter provides a description of the design process for the Eiffel Tower model using the SolidWorks software package. SolidWorks was developed by the French software company Dassault Systèmes, which also offers the CATIA software package—primarily intended for designing complex geometries in the aerospace industry. The graphical user interface of SolidWorks is tailored to a broad range of engineers, allowing for fast

learning and user-friendly operation. Unlike in the past, when conceptualizing ideas and constructing prototypes relied on sketches and technical drawings, today's 3D models allow for more thorough functionality checks. Combined with virtual simulations and environmental visualization, they enable easier decision-making and acceptance of developed prototypes. The development of 3D modeling software has brought numerous advantages: reduced design time, decreased workload for designers and engineers, improved quality of technical documentation, lower overall system design costs, and simplified unification of the entire workflow within a single project. As a result, the quality of an integrated project is significantly higher compared to working across different platforms—not to mention the limitations of manual drafting used in the past.

Figure 1a shows the Eiffel Tower model, with overall dimensions of 200×500 mm. Due to its complex lattice structure and hollow interior, it was necessary to reconstruct the model to make it manufacturable. Manufacturing the complete model using the FDM rapid prototyping method would require the use of support structures (Figure 1b), which are difficult to remove after printing and can negatively affect the dimensional accuracy and overall shape of the final product. In many cases, support removal requires the use of additional technologies, further increasing production costs and time. To avoid these undesired effects, the model was reconstructed by dividing the original geometry into nine simpler parts (Figure 1c). In addition, each segment was designed with appropriate technological features, allowing for the model to be assembled like LEGO blocks after manufacturing. This approach ensured the correct positioning and orientation of all components within the final assembly.

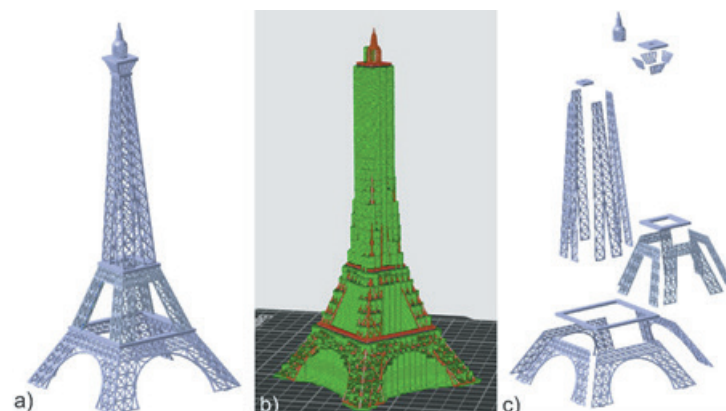


Figure 1. Eiffel Tower model: a) complete model, b) model with supports during manufacturing, c) exploded view showing all individual assembly components



3. RAPID PROTOTYPING TECHNOLOGIES

Rapid Prototyping refers to a layer-by-layer manufacturing technology used for producing parts. There are four main types of rapid prototyping technologies [1], schematically illustrated in Figure 2. What unifies them is the use of the standard .stl file format for preparing models for the selected manufacturing method. Whether the process involves material addition or subtraction, in addition to the .stl model of the desired prototype, an appropriate software tool is required for technology setup and generating machine instructions. The layer-by-layer material addition approach (Additive Manufacturing) is a natural solution for producing the presented Eiffel Tower prototype. This technology offers nearly unlimited design freedom, does not require special tooling, and can manufacture parts with mechanical properties comparable to those made by traditional manufacturing methods. Commonly referred to as 3D printing technologies, these methods have existed since the 1990s; however, their high cost and complexity at the time made them far less accessible than they are today. With technological advancement and cost reduction, 3D printers have become widely available. In-house 3D printing allows engineers and designers to quickly iterate between digital models and physical prototypes. Today, it is possible to manufacture a prototype within a single day and make multiple design modifications related to shape, size, or fit within the same timeframe, based on test results and real-world performance analyses.

Subtractive technologies represent a group of manufacturing methods in which material is removed layer by layer from the initial workpiece to achieve the desired shape and dimensions of the product.

Unlike additive technologies, which build up material, subtractive processes include milling, turning, drilling, grinding, and electrical discharge machining. These methods form the foundation of traditional manufacturing and are used across a wide range of industries, from automotive to precision engineering. Key advantages of subtractive technologies include high dimensional accuracy, excellent surface finish, and the ability to process a variety of materials, including metals, plastics, composites, and various types of model-making foams. An added benefit is the availability of user-friendly software for programming additive manufacturing machines, which typically allows for simple programming in just a few steps and does not require extensive user training. Hybrid techniques combine both additive and subtractive manufacturing processes. They are commonly used for producing functional parts, where the basic geometry is built using additive technologies, followed by precision machining of functional surfaces while the part is still positioned in the machine's working envelope. This approach enables high dimensional accuracy and superior surface quality, particularly after subtractive post-processing. Formative technologies are widely used for manufacturing parts through casting or forming processes. In practice, these methods are known as Rapid Tooling—the rapid production of tooling components. The tools are created using a combination of the three previously mentioned technologies, after which the desired material is poured into the prepared molds. Commonly used materials include epoxy-based resins for functional components and silicone-based compounds for creating molds intended for further use.

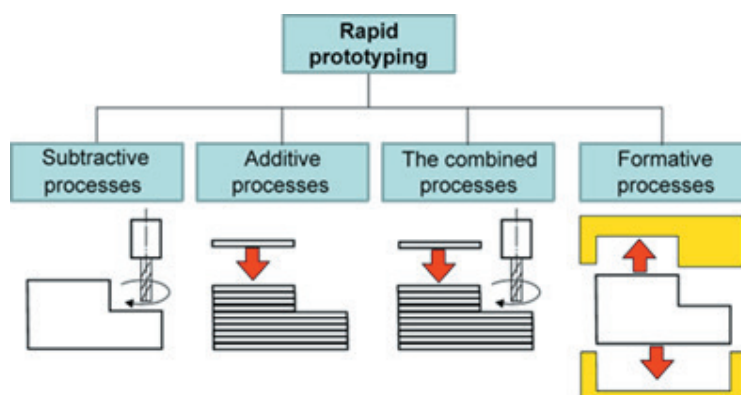


Figure 2. Overview of Rapid Prototyping Technologies [3]



4. PREPARING THE MODEL FOR FDM MANUFACTURING

This chapter presents the additive manufacturing machine used for producing the segments of the Eiffel Tower model and illustrates the preparation process of those segments. Finally, the manufactured parts were assembled into a complete structure, and the prototype model of the Eiffel Tower was presented.

4.1. MODEL OF THE BAMBU LAB P1S 3D PRINTER

When The mentioned 3D printer represents a new-generation additive manufacturing machine tool. It features an XY Core axis configuration, where the extruder moves within the XY plane while the platform gradually lowers along the Z-axis by the thickness of the defined layer. Movement in the XY plane is driven by a toothed belt, which offers several advantages. One of the main benefits is high speed and the possibility of implementing an H-bot mechanism. The Z-axis is powered by trapezoidal lead screws, which provide high accuracy and a self-locking effect. With the addition of the AMS (Automatic Material System) unit, which is purchased separately, the printer can produce parts in multiple colors. In essence, this unit functions as a material storage system that supplies the required filament during the printing process. The supporting structure of the Bambu Lab P1S 3D printer is a fully enclosed, box-type frame, which provides structural rigidity and maintains a stable ambient temperature—contributing to the high quality of the manufactured parts. Bambu Lab has developed its own proprietary software for programming all models in its 3D printer lineup. Additionally, the printer model used in this work is equipped with a touchscreen display for operator interaction and machine operation.

When connected to a network via Wi-Fi, the printer can receive programs directly and allows for real-time monitoring via a built-in camera and a mobile application. Figure 3 shows the Bambu Lab P1S 3D printer model.

4.2. PREPARATION OF THE PRINTING PROGRAM

This subsection provides a description of the procedure and software used for preparing the print program. The preparation was carried out using the Bambu Studio software package. In addition to basic model manipulation commands, the software offers a wide range of settings related to the printing process itself. It allows the user to define the number of layers on the top and bottom surfaces of the model, as well as the number of perimeters on the outer contours. Infill density is a critical parameter, as it directly affects the stiffness of the printed part, as well as its weight and material consumption. The software allows users to select the material type and, based on the selection, automatically recommends optimal extruder speeds and material deposition rates. The optimal printing speed of the extruder is 500 mm/s, with accelerations reaching up to 20,000 mm/s². Additionally, the P1S printer features advanced vibration compensation algorithms and improved pressure control during extrusion, which contribute to smooth and accurate prints even at high speeds. It is important to note that although the P1S can achieve high printing speeds, optimal results depend on the correct adjustment of print parameters and the use of appropriate materials [7]. Figure 4 shows a set of parts prepared for printing.

As part of the basic commands for model management, the user can position and orient parts on the build plate. Part orientation is especially important because the direction of layer deposition directly affects the mechanical strength of the part in different directions.

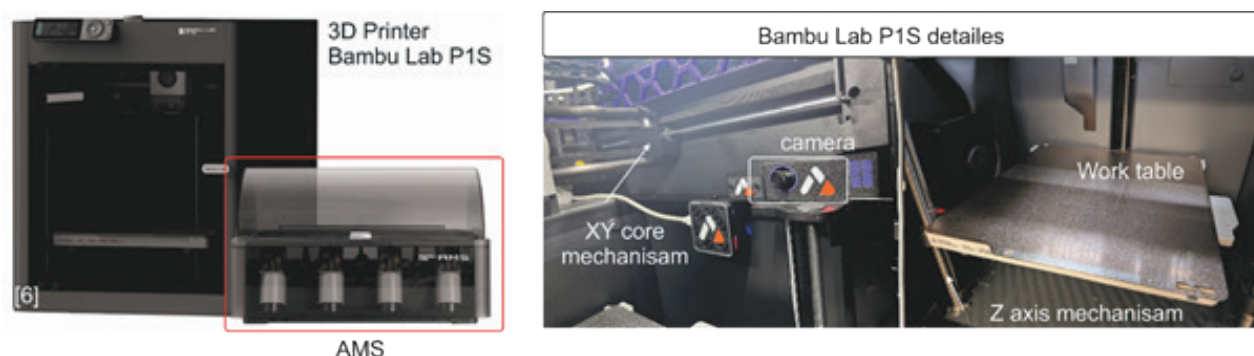


Figure 3. Bambu Lab P1S 3D Printer Model with Details



Irregularly shaped parts may require additional support structures, which are mechanically removed after printing—as shown in Figure 1b. Recently, support filaments based on cornstarch have been developed. These materials dissolve in water, making support removal much easier, especially for complex geometries like this one. The only requirement in this case is that the printer must be equipped with dual extruders, where one extrudes the support material and the other extrudes the main filament to build the part geometry. Parts prepared for printing can be scaled or duplicated if multiple copies are required. The software also includes an automatic arrangement function, which optimizes part placement on the build plate to maximize space utilization and enable simultaneous printing of as many parts as possible. This feature contributes significantly to reducing total manufacturing time. After completing the preparation of parts—which includes defining the printing parameters—it is necessary to slice the selected parts into layers of the specified thickness.

The software performs the necessary calculations and provides estimates of the total printing time and material consumption, which can be of great importance in certain situations. Rapid prototyping machine tools operate based on control systems that interpret instructions from a program written in G-code, defined by the ISO 6983 standard, also known as RS-274. Most G functions are standardized and identical across various types of 3D printers from different manufacturers. Differences between manufacturers may arise in the implementation of certain M functions, which control specific processes such as enabling extrusion, activating cooling fans, or turning on bed heaters via thermocouples. Figure 5 shows the assembled model of the Eiffel Tower. Thanks to the technological features added for positioning and orientation, the assembly process was completed quickly, and a specific two-component adhesive used to bond the parts provided additional structural strength to the final model.

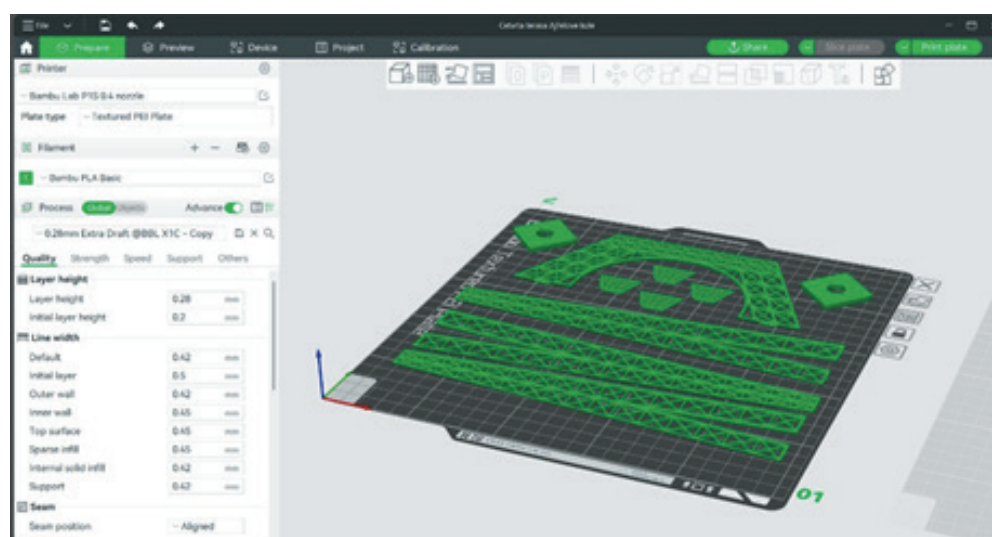


Figure 4. Bambu Studio Interface with Preparation of the First Group of Parts

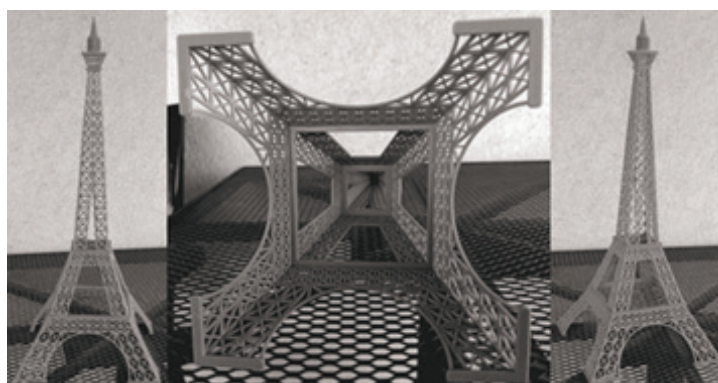


Figure 5. Assembled Model of the Eiffel Tower



5. CONCLUSION

The process of manufacturing the Eiffel Tower model using additive manufacturing proved to be an efficient and flexible method for producing parts with complex geometries. Thanks to the ability to directly convert CAD models into physical prototypes, it enabled rapid testing and visualization of intricate structures, as well as timely revisions of design solutions. Dividing the model into multiple segments further facilitated manufacturing, assembly, and quality control of the final structure. Although there are certain limitations regarding dimensional accuracy, printing speed, and the mechanical properties of the produced parts, the demonstrated method has proven to be highly suitable for creating models and visual prototypes. Rapid prototyping offers significant advantages to engineers and designers by accelerating product development, reducing costs, and shortening the time-to-market for new products.

REFERENCES

- [1] D. T. Pham and S. S. Dimov, *The Technologies and Applications of Rapid Prototyping and Rapid Tooling*, 1st ed. London, UK: Springer, 2000.
- [2] D. T. Pham and S. S. Dimov, "Rapid prototyping and rapid tooling—the key enablers for rapid manufacturing," *Proc. Inst. Mech. Eng. Part C: J. Mech. Eng. Sci.*, vol. 217, no. 1, pp. 1–23, 2003.
- [3] S. T. Zivanovic, M. D. Popovic, N. M. Vorkapic, M. D. Pjevic, and N. R. Slavkovic, "An overview of rapid prototyping technologies using subtractive, additive and formative processes," *FME Transactions*, vol. 48, no. 1, pp. 139–146, 2020.
- [4] M. B. Kumar and P. Sathiya, "Methods and materials for additive manufacturing: A critical review on advancements and challenges," *Thin-Walled Structures*, vol. 159, p. 107228, 2021.
- [5] T. G. Gawel, "Review of additive manufacturing methods," *Solid State Phenomena*, vol. 308, pp. 1–20, 2020.
- [6] Bambu Lab, "Official Website," [Online]. Available: <https://www.bambulab.com>. [Accessed: Mar. 27, 2025].
- [7] 3D Republika, "Recenzije, Vesti i Uputstva o 3D Štampi," [Online]. Available: <https://3drepublika.com>. [Accessed: Mar. 27, 2025].



APPLICATION OF ADDITIVE MANUFACTURING TECHNOLOGIES IN THE PRODUCTION OF A MONASTERY MODEL FOR THE PRESERVATION OF SERBIA'S CULTURAL HERITAGE

Dragan Cvetković^{1*},
[0000-0002-4405-3775]

Nikola Vorkapić²,
[0000-0001-5081-4828]

Ivana Cvetković²,
[0000-0002-2773-5643]

Aleksandar Cvetković³,
[0009-0003-3300-9845]

Srđan Trajković⁴
[0009-0005-1712-1065]

¹Singidunum University,
Belgrade, Serbia

²Faculty of Mechanical Engineering,
Belgrade, Serbia

³Jaroslav Černy Institute for
Water Management,
Belgrade, Serbia

⁴Academy of Applied Science Politehnika,
Belgrade, Serbia

Abstract:

In order to preserve Serbia's cultural heritage and monuments that are protected as cultural assets by the state, they have been digitized in a three-dimensional electronic form. This approach ensures the creation of a permanent digital record. Monasteries and monuments in digital form can serve as the basis for the development of avatars or models within Virtual Reality systems. On the other hand, such representations are suitable for various types of presentations and educational purposes, particularly for children and tourists who are not able to physically visit or experience these sites. Technological advancements have enabled the development of virtual environments that accurately replicate real-world spaces. This paper presents a methodology that includes modeling of the Sopocani Monastery based on photographs and original architectural drawings. After the virtual model was created, a physical 3D model of the monastery was produced using FDM (Fused Deposition Modeling) additive manufacturing technology. In addition, a mold was made from silicone material, which can potentially be used to produce replicas of the monastery using different materials. This mold was used to fabricate a model of the Sopocani Monastery in epoxy resin.

Keywords:

Rapid Prototyping, Additive Manufacturing, Rapid Tooling, 3D Printing.

INTRODUCTION

Cultural heritage represents the legacy passed down from previous generations and encompasses both tangible and intangible assets that are vital for the identity and continuity of a community. It includes archaeological sites, historical buildings, works of art, as well as traditions, language, and customs that shape a nation's cultural identity. Preserving and protecting these assets is essential for transmitting values, knowledge, and traditions to future generations. The protection of cultural heritage involves various measures aimed at identifying, documenting, researching, maintaining, and promoting cultural properties. These measures ensure sustainability and accessibility, preventing deterioration, damage, or destruction caused by natural disasters, urbanization, or neglect. Furthermore, protecting cultural heritage strengthens cultural identity, encourages cultural tourism, and positively impacts the economic development of communities.

Correspondence:

Dragan Cvetković

e-mail:

dcvetkovic@singidunum.ac.rs



In the modern age, technological advancements provide new opportunities for preserving and presenting cultural monuments. One innovative method is the creation of three-dimensional (3D) models of cultural assets. Developing 3D models allows for detailed digital documentation of objects, which is highly valuable for analysis, restoration, and education. The physical production of these models using 3D printers enables the creation of accurate replicas of monuments that can be used in museums, educational institutions, or for research purposes. These models facilitate interaction with cultural assets in ways that are not always possible at the original sites, thereby increasing accessibility and awareness of the significance of cultural heritage. One key technique for generating 3D models is photogrammetry. This process uses photographs of an object taken from different angles to create accurate three-dimensional models. Software packages such as Autodesk 123D Catch or Autodesk ReCap Pro convert these photographs into detailed 3D models, providing a valuable tool for documenting and analyzing an object's physical condition. This method is especially useful for documenting complex structures, allowing professionals to study and work on preserving cultural assets without requiring constant physical access to the location. The application of photogrammetric methods and 3D modeling significantly contributes to the protection and preservation of cultural heritage in Serbia. These research efforts provide valuable insights into the potential of modern technologies for documenting and restoring cultural monuments, thereby enhancing methodologies for protecting and promoting Serbia's cultural legacy. Figure 1 illustrates the application of the photogrammetric method on the monument of Commander Petar Bojovic.

Figure 1 illustrates the process of generating a 3D model of the bust of Petar Bojovic using the Autodesk 123 Catch software. This software works by uploading a series of photographs of the object, allowing it to reconstruct the environment and create a 3D model. The technique for capturing these photographs is crucial. The object should be photographed while moving in a circular path, with the camera positioned parallel to the surface on which the object stands, ensuring that the entire object is recorded. After completing the first circular pass, a second round of photographs should be taken at a 45-degree angle relative to the base, again ensuring comprehensive coverage of the object. Each pass should consist of at least 10 to 15 images. Photographs that display glare or reflections are not useful and may compromise the quality of the 3D model reconstruction. Therefore, it is advisable to take photos under cloudy conditions and to apply a matte coating to reflective surfaces if necessary. Once the set of photos is uploaded, a 3D environment is generated, which can be cleaned and edited using built-in tools to remove any unwanted elements. In the final step, a 3D model in STL format is created, making it suitable for use with any rapid prototyping software. As shown in Figure 1c, a prismatic base has been added to the lower part of the model to serve as a stand for the bust. This method is also the most cost-effective technique for 3D model reconstruction. The second technique for generating 3D objects is simpler to perform but significantly more expensive due to the necessary equipment. It involves the use of a portable handheld 3D scanner, paired with a powerful computer. This method yields a more accurate model in terms of shape and dimensions since the scanner generates a dense point cloud connected by triangles that form the external contour of the object. The output file is also in STL format, ready for direct use on additive manufacturing machines.

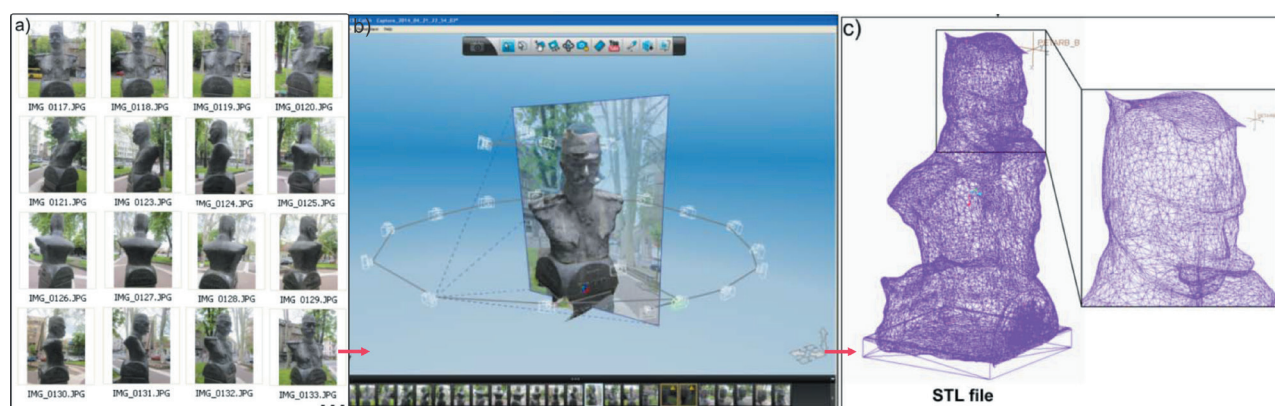


Figure 1. Photogrammetric method applied to the bust of Petar Bojovic, adapted from [3]



The third technique is the simplest and least expensive but is also the most time-consuming. It involves manual modeling using one of the available software packages, based on sketches and measurements taken on-site. This method can be seen as a form of reverse engineering of the analyzed object, allowing for a model that can be easily refined and converted into any required format for further applications. One potential use of such models is in virtual reality environments or for developing video game assets. This first chapter provides an overview of the available techniques for generating 3D objects. The second chapter will describe the specific object that will be further examined. The third chapter outlines the mold manufacturing process for producing a scale model of the object, while the fourth chapter discusses the applications of the completed mold.

2. MANUFACTURING OF A 3D MODEL OF THE SOPOCANI MONASTERY

This chapter offers an overview of the Sopocani Monastery, featuring a 3D model created using SolidWorks software. The Sopocani Monastery was constructed by King Stefan Uros (1243–1276) near the source of the

Raska River, located approximately 17 km from the city of Novi Pazar. As a significant site within the medieval Serbian state of the Nemanjic dynasty, King Stefan the First-Crowned's third son left behind a legacy that, in both scale and beauty, surpassed all previous Serbian churches. The frescoes of the Sopocani Monastery are true masterpieces of artistry, and combined with its stunning architecture, they have garnered international recognition for the site. Today, the Sopocani Monastery is regarded as one of the most important cultural monuments in Serbia and has been included in the UNESCO World Heritage List since 1979, as part of the protected medieval ensemble "Stari Ras and Sopocani." [4] Figure 2 displays photographs of the Sopocani Monastery.

The photographs reveal distinct decorative elements and details along the edges of the windows and arches. While these features are quite small compared to the overall size of the monastery, they are significant. The monastery was designed to full scale, based on the documentation presented in Figure 3, which includes a properly scaled architectural drawing of the structure. A potential issue arises when manufacturing a scaled-down model of the monastery, which is 30 times smaller than the original. In this case, certain details will also



Figure 2. Photographs of the Sopocani Monastery

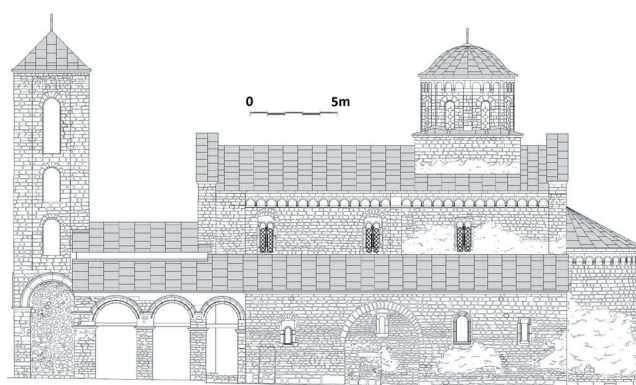


Figure 3. Documentacion of the Sopocani Monastery



be reduced by a factor of 30, potentially making them too small to be accurately reproduced. The solution is to refine the 3D model further, intentionally enlarging specific dimensions to ensure they can be effectively manufactured.

Figure 4 shows a scaled-down 3D model of the Sopocani Monastery at a scale of 30:1. It can be observed from the image that certain details, such as individual bricks or roof tiles, are not visible, simply because their dimensions affect the manufacturability of the model. Since the model of the Sopocani Monastery was manufactured using FDM (Fused Deposition Modeling) rapid prototyping technology, the gaps between bricks or the size of the tiles would influence the layer thickness during printing, and such small features would negatively impact the surface quality of the final model.

A more detailed version of the monastery model, including fine architectural elements, would be better suited for manufacturing using SLA (Stereolithography) technology a layer-by-layer additive process that is one of the oldest rapid prototyping methods. This technique is based on curing liquid polymer using a UV laser.

3. MANUFACTURING OF THE MONASTERY MODEL USING RAPID PROTOTYPING TECHNOLOGIES

This chapter will present two different available rapid prototyping technologies using the Sopocani Monastery as a case study. The first technology describes the manufacturing of the Sopocani Monastery model using FDM (Fused Deposition Modeling) rapid prototyping technology. The second technology represents the rapid production of casting tools (Rapid Tooling), which essentially uses rapid prototyping methods and silicone mixtures, rather than commercial technologies and materials, to create tools into which different materials can later be cast.

3.1. MANUFACTURING OF THE MONASTERY MODEL USING FDM TECHNOLOGY

The prepared STL file was imported into Bambu Studio software, where all key printing parameters were defined, including layer thickness, infill density, nozzle, and heated bed temperatures, as well as support structure generation.

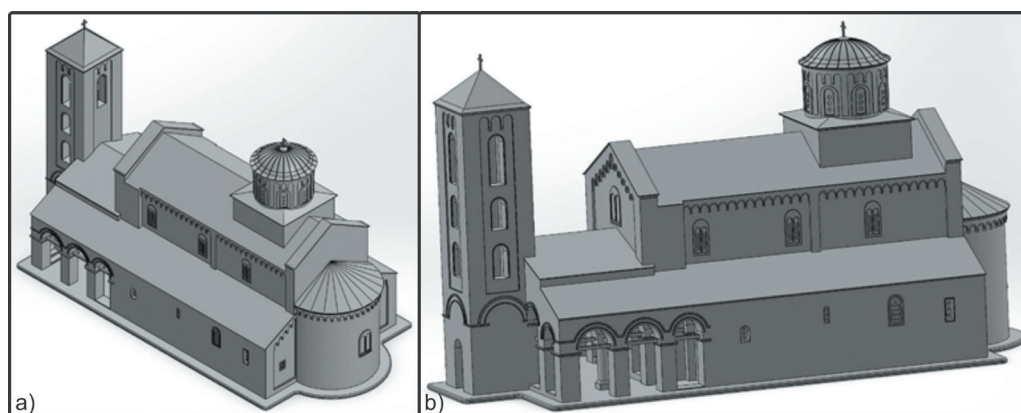


Figure 4. 3D Model of the Sopocani Monastery Created in SolidWorks Software Package

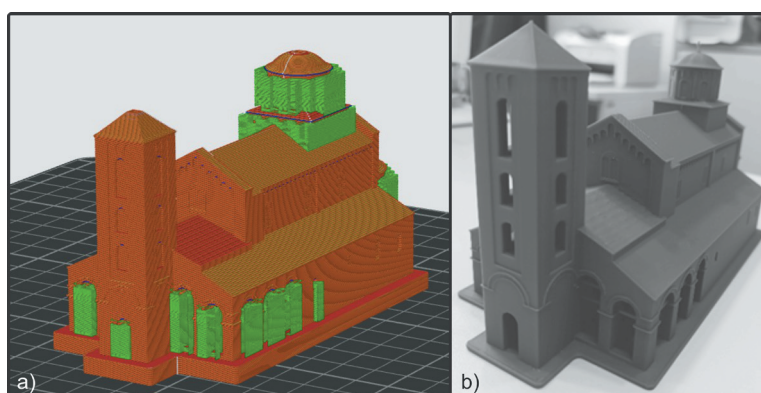


Figure 5. 3D Printed Model of the Sopocani Monastery



The model was then manufactured using the Bambu Lab P1S 3D printer, known for its high speed and precision, as well as its ability to reliably reproduce complex geometries. Figure 5a shows the preparation of the monastery model in Bambu Studio, while Figure 5b presents the printed model of the Sopocani Monastery. Upon completion of the printing process, minimal post-processing was required, consisting solely of the removal of support structures generated during printing. The support material is shown in green in Figure 5a. Thanks to the printer's precision and well-optimized parameters in Bambu Studio, additional procedures such as sanding, gluing, or painting were not necessary, which simplified and accelerated the entire manufacturing process.

3.2. MANUFACTURING OF THE MONASTERY MODEL USING RAPID TOOLING TECHNOLOGY

Rapid tooling refers to manufacturing techniques that leverage rapid prototyping technologies. In this study, rapid tooling is specifically used to create a mold for casting the model of the Sopocani Monastery. This concept involves producing a mold using a deformable silicone material, which allows for the accurate transfer

of intricate details from the positive model to the negative mold. Certain geometric features of the monastery model exhibit what is known as negative geometry. If the mold were constructed from rigid materials such as metals, plaster, or sand, these features could not be removed without damaging the mold. The positive model of the Sopocani Monastery can be produced via either subtractive or additive manufacturing technologies. Since the model has already been created using Fused Deposition Modeling (FDM) technology, it is necessary to design a two-part, detachable mold that can accommodate the silicone mixture.

Figure 6 illustrates the step-by-step process of forming the silicone mold. The box that holds the monastery model was also manufactured using FDM technology and designed for easy assembly and disassembly, as shown in Figures 6a and 6b. The silicone compound used in this process is called 5.0 Premium, which consists of two components: a base material and a curing agent, mixed in a 1:1 ratio before being poured into the mold containing the positive model of the Sopocani Monastery. After one hour of curing, the outer box mold can be removed from the formed silicone mold. The final silicone mold is displayed in Figure 6c.

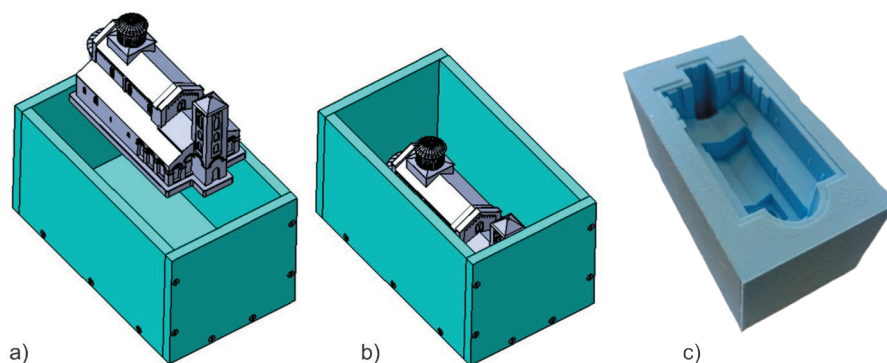


Figure 6. Silicone Mold Fabrication Process

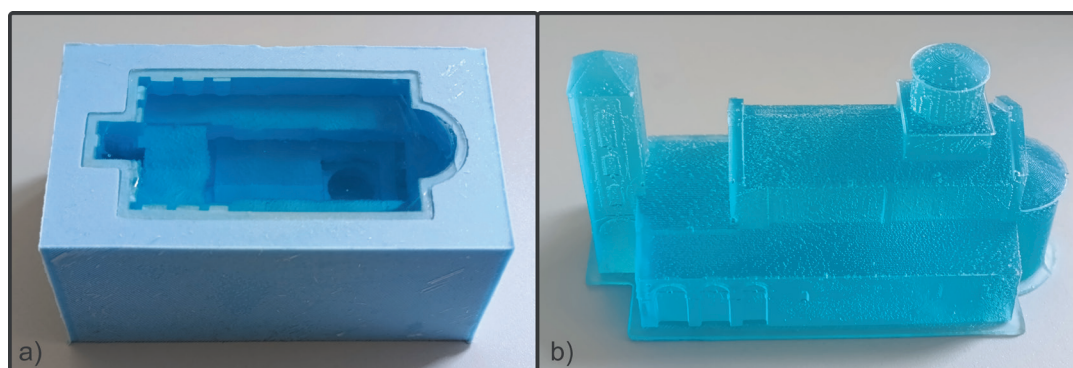


Figure 7. Process of Manufacturing the Sopocani Monastery Model Using Epoxy Resin



The resulting mold is a finished product ready for use. Various types of liquid polymers, such as two-component epoxy resin, can be poured into the mold. This epoxy resin hardens after a specific period of time. The initial curing period is 24 hours after pouring, but the material reaches its final properties after 10 days of curing. Since epoxy resin is a two-component system mixed with a hardener in a 2:1 ratio, it is essential to use a heat gun or torch during the pouring process. This is done to treat the upper surface of the cast material and eliminate any air bubbles formed during mixing. Once poured into the mold, the epoxy resin generates a certain amount of heat, which helps expel bubbles from the mixture. Figure 7a shows the epoxy resin poured into the silicone mold, while Figure 7b displays the finished scale model of the Sopocani Monastery made from epoxy resin.

4. CONCLUSION

The use of additive manufacturing technologies for preserving cultural heritage represents an innovative blend of modern engineering and cultural stewardship. This is exemplified by the production of a scale model of the Sopocani Monastery, which showcases the effectiveness of Fused Deposition Modeling (FDM) technology in creating physical representations of historically significant structures. The monastery was digitized using SolidWorks software, and then a 3D model was printed with a Bambu Lab P1S printer. This process allowed for the accurate and rapid production of a scaled-down version of the monastery, requiring minimal post-processing. Additionally, a silicone mold was created to facilitate the serial production of models, which can be made from various materials, such as epoxy resin.

This approach not only aids in the preservation of cultural heritage through documentation and education but also opens up opportunities for museum exhibits, souvenirs, and interactive educational content. The combination of digital technologies and additive manufacturing offers a wide array of applications in protecting, promoting, and educating about cultural and historical heritage, enhancing the accessibility, durability, and appeal of cultural content.

REFERENCES

- [1] D. Đukic, "Zaštita kulturnih dobara i imovinska prava crkava i verskih zajednica," *Harmonius Journal of Legal and Social Studies in Southeast Europe*, vol. 4, pp. 80–98, 2020. [Online]. Available: <https://www.harmonius.org/wp-content/uploads/2021/04/Pages-from-Harmonius-2020-06.pdf>. [Accessed: Apr. 1, 2025].
- [2] ICCROM, *First Aid to Cultural Heritage in Times of Crisis: Handbook for Coordinators*. Rome: International Centre for the Study of the Preservation and Restoration of Cultural Property, 2018. [Online]. Available: https://www.iccrom.org/sites/default/files/publications/2024-02/fac_handbook_digitalprint_oct-2018_final_ser.pdf. [Accessed: Apr. 1, 2025].
- [3] S. Živanovic, S. Tabakovic, and S. Randjelovic, "Rapid prototyping of art sculptural shapes according to the sample," **Advanced Technology Material*, vol. 44, no. 1, pp. 27–32, 2019.
- [4] UNESCO World Heritage Centre, "Stari Ras and Sopocani," UNESCO World Heritage Centre. [Online]. Available: <https://whc.unesco.org/en/list/96/>. [Accessed: Apr. 1, 2025].
- [5] C. K. Chua, K. F. Leong, and Z. Liu, "Rapid Tooling in Manufacturing," in *Handbook of Manufacturing Engineering and Technology*, A. Nee, Ed. London: Springer, 2013. [Online]. Available: https://doi.org/10.1007/978-1-4471-4976-7_39-1
- [6] S. T. Zivanovic, M. D. Popovic, N. M. Vorkapic, M. D. Pjevic, and N. R. Slavkovic, "An overview of rapid prototyping technologies using subtractive, additive and formative processes," *FME Transactions*, vol. 48, no. 1, pp. 139–146, 2020.



OPEN-SOURCE AND CLOUD-BASED SOLUTIONS FOR EFFICIENT HFSWR SITE PLANNING

Darko Marjanović^{1,2},
[0000-0002-7731-8645]

Dragan Golubović^{2*}
[0000-0003-0019-0417]

¹Singidunum University,
Belgrade, Serbia

²Vlatacom Institute,
Belgrade, Serbia

Abstract:

The capability to detect targets beyond the horizon is one of the High Frequency Surface Wave Radar's (HFSWR) key advantages, which are primarily used for detecting vessels at distances exceeding 350 km. Compared to microwave radars, which typically have detection ranges of up to 40 km, HFSWRs offer significantly greater coverage. This paper specifically focuses on planning the coverage area of these radars using the proposed solution and open-source software tools. The proposed software tool should streamline marketing presentations, site planning, and site surveys by enabling fast and efficient coverage planning, the selection of HFSWR locations, coverage areas plotting, and the visualisation of antennas/containers, directly on Google Earth (GE). GE is a freely available desktop and mobile application which enables preliminary site visualization where HFSWR is planned. We also identified readily available, widely adopted, and free applications (open-source) that provide excellent results with minimal programming effort. Although the scripting frameworks of licenced software tools share structural similarities, the functional workflow development often demands significant time to identify suitable functions and write new code, but they are cost-effective (Google Sheets, LibreOffice Calc etc.), free and widely accessible, with no licensing costs. Cross-platform flexibility and cloud-based mode of operation are of particular importance, due to the possibility of calculating coverage area using only mobile devices, which is especially important in practice. The rapid visualization enables us to adjust input parameters and redraw multiple times until we achieve an optimal solution for each HFSWR location, as well as comprehensive coverage across multiple locations.

Keywords:

HFSWR, OTHR, Radar Location Planning, Antenna Positioning, Antenna Arrays.

INTRODUCTION

Radars are commonly used to detect remote objects using electromagnetic waves. However, High Frequency Surface Wave Radars (HFSWRs) differ significantly from conventional radars. Unlike conventional radars, HFSWR is neither mounted on towers nor compact enough to fit in vehicles or aircraft. Additionally, a direct line of sight between the radar and the target is not required for target detection. This capability to detect targets beyond the horizon is one of HFSWR's key advantages.

As the name suggests, HFSWR operates in the HF band (3–30 MHz). In this frequency band, a surface wave component forms between the electromagnetic wave and the sea surface.

Correspondence:

Dragan Golubović

e-mail:

dragan.golubovic@vlatacom.com





This surface wave follows the Earth's curvature, enabling reflections from targets far beyond the line of sight. Given the long wavelengths (10–100 m) in the HF band, large reflective surfaces are required. Consequently, HF-SWR is primarily used for detecting vessels beyond the horizon at distances exceeding 350 kilometres. Compared to microwave radars, which typically have ranges of up to 40 km, HFSWR offers significantly greater coverage. The hardware basis of HFSWRs, as well as data on the antenna arrays used in such systems, can be found in [1] [2] [3] [4].

The primary market for HFSWR systems includes agencies responsible for a country's Exclusive Economic Zone (EEZ). Preventing illegal fishing, drug smuggling, illegal immigration, theft of crude oil, terrorism and support for rescue missions can return investment in such a kind of investment very fast. Unlike satellite or aerial surveillance, HFSWRs provide continuous, cost-effective monitoring. Many authors deal with HFSWR system design and radar signal processing, as presented in [5] [6]. They also describe how these radars actually work in practice, what their coverage areas are, system parameters, and generally what kind of signal processing leads to good results when it comes to tracking of targets, or ships in this case. In order to be able to simulate the coverage area in general, it is very important to have a good knowledge of all parts of these systems.

HFSWR allows the surveillance of wide zones at large distances. Well-planned coverage actually means that we can detect targets in a large spatial sector with no or very few blind spots, especially when it comes to a radar network, where it is necessary to use multiple radars to cover a large length of coastline. The next step relates to signal processing, which should result in the detection of all targets of interest in a spatial sector. In recent years, high-resolution primary signal processing has been of particular interest, due to its numerous advantages over classical signal processing [7] [8] [9].

HFSWR system design and performance analysis require a numerical tool capable of representing complex environments to model wave propagation in realistic conditions. Many authors deal with modelling of HFSWRs, as presented in [10] [11] [12] [13]. These papers actually form the basis for the development of the coverage area simulator presented in this paper.

The proposed software tool should streamline marketing presentations, site planning, and site surveys by enabling the visualisation of antennas/containers and coverage zone, directly on Google Earth (GE). They are cost-effective, such as Google Sheets and LibreOffice Calc, free and widely accessible, with no licensing costs. Cross-platform flexibility and cloud-based mode

of operation are of particular importance, due to the possibility of calculating coverage area using only mobile devices.

In practice, after such a simulation, where the position of all elements of the HFSWR system can be quickly determined, the actual design of such a system follows, and finally the implementation of algorithms for primary signal processing, as presented in [14] [15] [16].

The paper is organized as follows. In Section 2, the main principles of HFSWR site planning are presented. In Section 3, we present open-source and cloud-based software tools for efficient HFSWR site planning. In Section 4, the simulation results, based on an open-source solution for HFSWR coverage, are shown, while in Section 5 we make some conclusions.

2. HFSWR SITE PLANNING FUNDAMENTALS

Planning the coverage area is the first step in designing such systems. Frequency Modulated Continuous Wave (FMCW) is commonly employed in HFSWR systems. To mitigate challenges such as Bragg scattering and ionospheric interference, a dual-frequency operation is advantageous. Therefore, two antenna arrays are utilized on both the Rx and Tx sides. In practice, the receiver (Rx) area typically spans approximately 500 meters along the coastline and extends up to 100 meters inland, while the transmitter (Tx) area dimensions are roughly 100×100 meters. These dimensions depend on the operational frequencies. Typical HFSWR system architecture is shown in Figure 1.

A linear monopole antenna array provides 120° azimuth coverage, with an optimal spacing of $0.45 \cdot \lambda$ between antenna elements (where λ is the wavelength). In practice, an array of 32 monopole antennas achieves sufficient sensitivity for weak signals and high angular resolution. However, this configuration requires an extended coastal area at low elevation. The azimuth coverage aligns with the coastline geometry, ensuring the monitored area encompasses the country's EEZ.

To ensure effective coupling between transmitting/receiving antennas and the water surface, antennas must be installed near water at low elevations. Due to the extremely weak reflected signals, the receiving array requires a long antenna configuration. Tx power is crucial for long-range detection, requiring sufficient physical separation between the Tx and Rx arrays to prevent Rx saturation caused by its own signal.

The Tx array must be directional, focusing energy toward the coverage area while minimizing radiation toward the Rx array.

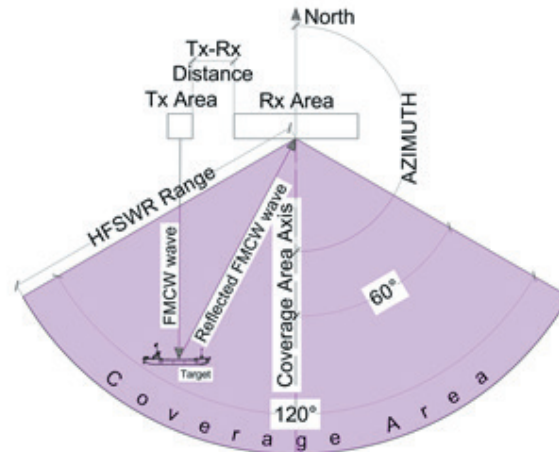


Figure 1. Typical HFSWR system architecture

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	
1	Type .kml filename:				EcuadorHFSWR_DemoLibOff_2										START								
2	Num	Name	Longitude	Latitude	OTHR Radius (km)	Azimuth (°)	Description	LF Freq (MHz)	HF Freq (MHz)	Tx-Rx Distance(n°A)	Dead Zone HFSWR (km)	Tx offset	Number of Antennas per Frequency	Color	Draw Order	Transparency [0-255]							
3	1	Isla Puná	-80.266195	-2.982152	350	280	OTHR	8.55	16.5	55	1	200	32		2	150							
4	2	Carrizal	-80.034123	0.288233	350	269	OTHR	8.55	16.5	70.15	1	-60	32		2	150							
5	3	Campo Alegre	-79.796494	0.904702	350	317	OTHR	8.55	16.5	59	1	-350	32		2	150							
6	4																						

Figure 2. LibreOffice input sheet

EcuadorHFSWRDemo																					
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Enter filename (.kml)				HFSWR_DemoOffset							☐	START CHECKBOX	START							
2	Num	Name	Longitude	Latitude	OTHR Radius (km)	Azimuth (°)	Description	LF Freq (MHz)	HF Freq (MHz)	Tx-Rx Distance (n°A)	Dead Zone OTHR (km)	Tx offset	Number of Rx Antennas per freq.	Color	Draw Order	Transparency [0-255]	HF Freq (MHz)				
3	1	Location1	-80.266195	-2.982152	350	280	OTHR	8.55	16.5	30	1	200	32		2	150	16.5				
4	2	Location2	-80.034123	0.288233	350	269	OTHR	8.55	16.5	70.15	1	-60	32		2	150	16.5				
5	3	Location3	-79.796494	0.904702	350	317	OTHR	8.55	16.5	59	1	-350	32		2	150	16.5				
6	4																				

Figure 3. Google Sheets input sheet

Planning the coverage area is very important, so it is essential to develop a software tool to do this step in the fastest way. In the next section, one such solution will be presented, with a focus on open-source solutions.

3. OPEN-SOURCE AND CLOUD-BASED SOFTWARE TOOLS FOR EFFICIENT HFSWR SITE PLANNING

The topic of this section is the introduction of a software tool for HFSWR coverage estimation. This rapid visualization allows us to adjust input parameters and redraw them multiple times until we achieve an optimal solution for each HFSWR site. Also, a comprehensive EEZ coverage across multiple locations can be shown.

The generated drawing serves as both a marketing tool for potential customers and a valuable resource for preliminary site planning.

The software tool generates .kml files (a format natively supported by Google Earth). Input data processing, .kml file creation, and altitude extraction were automated using Microsoft Office Excel VBA. But in this paper, we present two open-source solutions by using LibreOffice Calc VBA and Google Sheets (for cloud-based collaboration). These platforms serve as user-friendly interfaces for data entry, reporting, and scripting (via VBA/JavaScript).

The input data sheet interface in LibreOffice Calc closely mirrors Microsoft Excel, ensuring compatibility in layout, formulas, and scripting workflows (e.g., VBA).



Microsoft Office uses VBA (Visual Basic for Applications) for scripting, while LibreOffice employs its own API-based Basic code (LibreOffice Basic). Though both languages share a similar syntax, their object models and methods differ significantly, as shown in Listing 1 and Listing 2.

Notably, VBA scripting is not supported in browser-based editions. So, Desktop users can initiate the script via a dedicated START button or a checkbox in the Google Sheets interface, as presented in Figure 2 and Figure 3. The structure of input sheets is shown in these figures.

START is a control button which can start the script from Sheet2. The script generates a .kml file that automatically opens in Google Earth within seconds and

enables the coverage area visualisation and antenna/container positions. The associated spreadsheets output the antenna coordinates (latitude, longitude) and altitudes as well. The generated .kml file is automatically saved in the same directory as the source spreadsheet file from which the script is executed. Table 1 shows all necessary input parameters and their detailed description. By using the LF array central position, antenna array angle, and inter-antenna distance, all antenna positions are determined. KML files outlining the EEZ of individual countries are also available online. By integrating these EEZ boundaries into custom .kml file, it is possible to enhance the visualization, enabling a clearer assessment of HFSWR coverage.

Table 1. Key hardware and software components of AI hearing aids

Parameter	Description
Name	Location name
Latitude and longitude	Array centre coordinates of the LF antenna array
Range (km)	Maximum coverage distance
Azimuth (degrees)	The orientation of the coverage area axis related to the north
Description	This section is reserved for future sensor type selection, such as the integration of microwave (MW) radars or cameras into the same script framework.
HF and LF Freq (MHz)	In this context, "LF" refers to a lower frequency within the HF frequency range, while "HF" denotes a higher frequency relative to another reference frequency
Tx-Rx distance	The separation between the Tx and Rx arrays is defined as a multiple of the lower frequency (LF) wavelength (λ).
Dead zone OTHR [km]	IT refers to the circular area near the radar system where targets cannot be detected. This occurs because the distance between the target and the radar must be significantly greater than both the operational wavelength and the Tx-Rx distance.
Tx offset	The Tx Offset quantifies the misalignment between the Rx and Tx arrays, defined as the horizontal distance of the Tx array from the Rx array's central axis (ideally minimized to zero).
Number of Rx antennas per Frequency	The number of receiving antennas allocated to each frequency band is typically between 16 and 32
Colour	The coverage area of the HFSWR is represented as a polygon. Users can customize the fill colour for each coverage area by selecting the desired fill colour in the corresponding cell. This feature enhances clarity when analysing overlapping zones or presenting results to stakeholders.
Draw order	Layering definition for overlapping polygons
Transparency	Opacity settings of the HFSWR coverage area polygon.

```
' Input Lat1 and Lon1 [decimal degrees] are starting coordinates, dist - distance between starting and
calculating coordinates [km],
' Azimuth - calculating from starting toward calculating coordinates [deg]
' Output: Lat2 and Lon2
Sub LatLonDist(ByVal Lat1 As Double, ByVal Lon1 As Double, ByVal dist As Double, ByVal Azimuth As Double,
ByRef Lat2 As Double, ByRef Lon2 As Double)

Dim Pi As Double
Pi = 3.14159265358979
r = 6378.1 ' km - radius of the Earth

Lat2 = WorksheetFunction.Asin(Sin(Lat1 * Pi / 180) * Cos(dist / r) + Cos(Lat1 * Pi / 180) * Sin(dist
/ r) * Cos(Azimuth * Pi / 180)) * 180 / Pi

Lon2 = Lon1 + WorksheetFunction.Atan2(Cos(dist / r) - Sin(Lat1 * Pi / 180) * Sin(Lat2 * Pi / 180),
Sin(Azimuth * Pi / 180) * Sin(dist / r) * Cos(Lat1 * Pi / 180)) * 180 / Pi

End Sub
```

Listing 1. An example of a program written in VBA Script (LatLonDist function)



```
function LatLonDist(lat1, lon1, dist, azimuth) {
  var pi = Math.PI;
  var r = 6378.1; // km - radius of Earth

  var lat1Rad = lat1 * pi / 180;
  var lon1Rad = lon1 * pi / 180;
  var azRad = azimuth * pi / 180;

  var lat2Rad = Math.asin(Math.sin(lat1Rad) * Math.cos(dist / r) + Math.cos(lat1Rad) * Math.sin(dist / r)
  * Math.cos(azRad));
  var lon2Rad = lon1Rad + Math.atan2(Math.sin(azRad) * Math.sin(dist / r) * Math.cos(lat1Rad), Math.
  cos(dist / r) - Math.sin(lat1Rad) * Math.sin(lat2Rad));
  return [lat2Rad * 180 / pi, lon2Rad * 180 / pi];
}
```

Listing 2. An example of a program written in JavaScript (LatLonDist function)



Figure 4. Rx antenna array (first 3 parts) and Tx antenna array (last part) which are displayed on a cell phone

4. SIMULATION RESULTS

The simulation results are generated for an arbitrarily chosen location in Ecuador, where complete coverage of the seashore is required. For this purpose, the coverage will first be shown by using a single radar, and then by using a radar network.

In case of using Google Sheets, the .kmz file will be saved automatically to Google Drive for immediate access to Google Earth. In cases where Google Earth fails to load the .kml file directly from Google Drive on mobile devices, the file must first be downloaded to the device's local storage and opened manually.

The script comprises four core functions. The primary function, *HFSWR_position()*, is initiated by a START button. During its execution, it calls two helper functions—*LatLonDist()* and *GetAltitudeD()*—as needed. Additionally, the *onEditTrigger(e)* function monitors user interactions, and automatically launches the main process when a designated checkbox is activated. The *GetAltitudeD()* function retrieves altitude data for

a known antenna location via the Bing Maps API and stores this information in the corresponding cell of Sheet 3. Note that this function will need to be updated in the future, as Microsoft has announced the retirement of the Bing Maps API. Therefore, the migration to Azure Maps is recommended for continued functionality. The *LatLonDist()* function calculates and returns latitude and longitude values.

Figure 4 shows Tx and Rx antenna positions which are displayed on a cell phone by reading output .kmz file from Google Drive. The user's real-time position is displayed as a blue marker on Google Earth. This feature aids in verifying antenna alignment, optimizing their placement, or ensuring safety protocols during site inspections.

The distance between the LF Rx array and the HF Rx array is equal to λ_{LF} , and the bearing is the azimuth. Using this information, we compute the central point of the HF Rx array and subsequently determine all HF Rx antenna positions. The container location with electrical equipment is also calculated to maintain a safe distance.



Ideally, the centre of the Tx antenna array should be aligned with the corresponding Rx array axis. All calculated antenna positions are stored in Sheet4 and they are categorized by location name and antenna array number. For the antenna positions, we illustrate lines representing antenna monopoles and radials. The height of each antenna is determined as follows: $H_a=\lambda/8+1.5\text{m}$, where $\lambda/8$ is the monopole height, and 1.5m is the bracket height. The Tx area is calculated by using the same Lat/Lon formula, with the Tx-Rx distance provided in the input sheet (Sheet 2). Tx antennas are arranged in a rectangular configuration with dimensions of $\lambda/2$ and $\lambda/4$, where $\lambda/2$ is parallel to the shoreline and the Rx array line of the corresponding frequency. Users can adjust parameters and regenerate site layouts within seconds. For example, if an antenna array's proximity to water risks flooding, its latitude and longitude can be updated directly in the spreadsheet.

The simulated coverage area by one HFSWR (cell phone preview) and corresponding antenna positions are shown in Figure 5, while Table 2 shows the corresponding altitudes for first 12 Rx antennas for 3 different HFSWR locations. The main goal is to cover the entire sea coast, which is impossible with just one radar.

For on-site visits, the .kml file can be downloaded onto a mobile phone, providing real-time orientation. During a site survey, engineers can access precise information about the locations of antennas and containers in their vicinity. If the designer identifies necessary changes during the site survey, a Google Sheets script is available for final adjustments. The updated coordinates can then be forwarded to local surveyors to identify property owners and facilitate ownership agreements between the system user and property owners. Figure 6 shows the simulation of coverage of the entire sea coast by using a radar network.

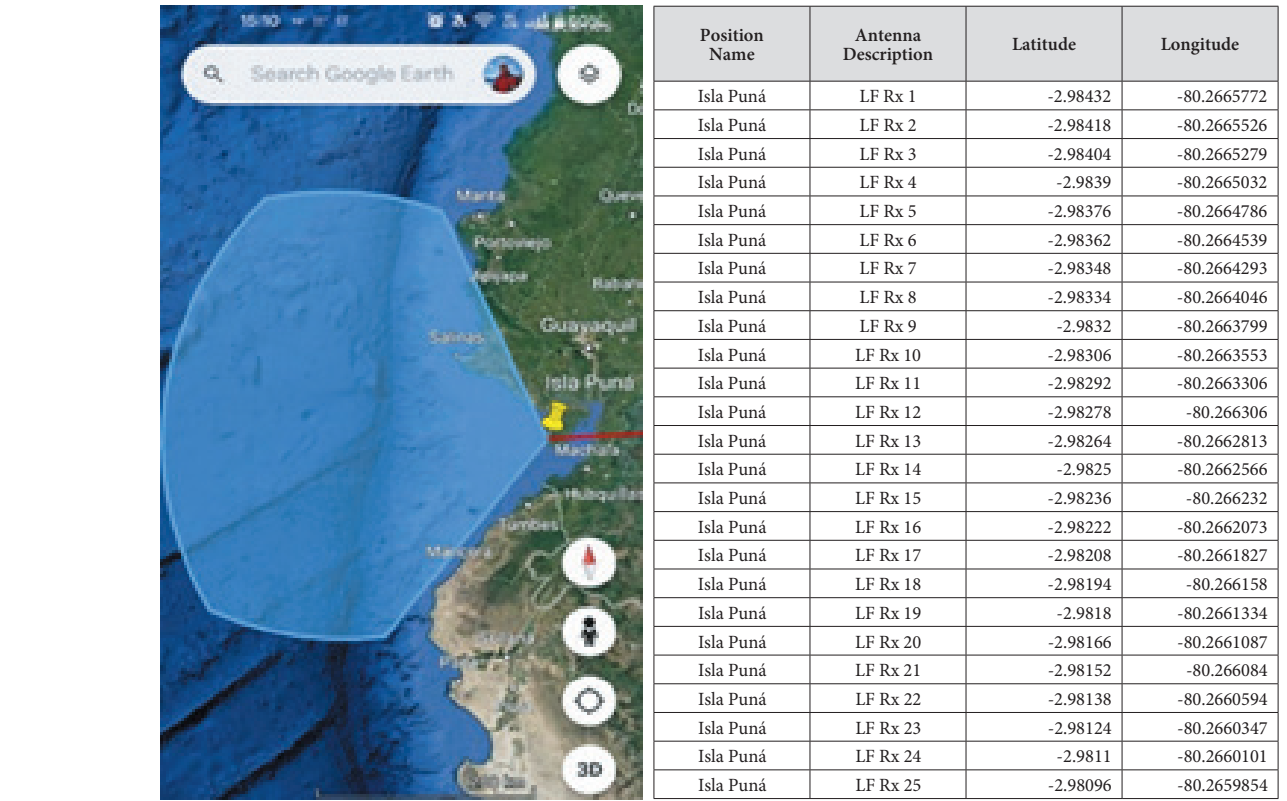


Figure 5. The simulated coverage area by one HFSWR (left), and corresponding antenna positions (right)

Table 2. The corresponding altitudes for first 12 Rx antennas

Locations/Sites	LFRx1	LFRx2	LFRx3	LFRx4	LFRx5	LFRx6	LFRx7	LFRx8	LFRx9	LFRx10	LFRx11	LFRx12
Isla Puná	1	1	1	1	1	1	1	1	1	1	1	1
Carrizal	6	6	6	6	6	6	6	6	6	6	6	6
Campo Alegre	9	9	9	10	10	10	10	10	10	10	11	11

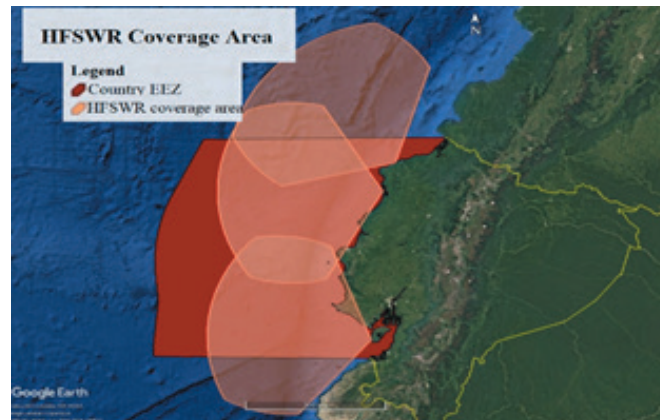


Figure 6. The coverage of the entire sea coast by using a radar network

5. CONCLUSION

In this paper, we proposed the open source and cloud-based software solution for HFSWR site planning that enables fast and efficient coverage planning, coverage areas plotting and the visualisation of hardware elements (antennas/containers), directly on Google Earth. Of particular value is the ability to display the coverage area of a multiple radar network, making it possible to easily simulate coverage of the entire sea coast, which is impossible to cover with just one radar. The key advantages of this solution include cost efficiency and cross-platform flexibility. We use Google Sheets and LibreOffice Calc which are widely accessible, with no licensing costs. These platforms serve as user-friendly interfaces for data entry, reporting, and scripting (via VBA/JavaScript). Another advantage is rapid iteration, because users can adjust parameters and regenerate site layouts within seconds. The proposed software tool shows the entire site in Google Earth, ensuring immediate validation of the required coverage area. Of particular importance is the application on mobile devices, especially during site surveys. Related to the main drawbacks, although macro-based workflows pose security risks in general, they are acceptable here due to internal file-sharing practices and special purposes. Also, LibreOffice, comparable to Microsoft Excel, has a smaller support community and fewer specialized resources, but JavaScript has extensive developer support and enables real-time scripting on mobile devices.

6. ACKNOWLEDGEMENTS

We thank Vlatacom Institute for overall supporting this research.

REFERENCES

- [1] D. Trizna and L. Xu, "Target Classification and Remote Sensing of Ocean Current Shear Using a Dual-Use Multifrequency HF Radar," *IEEE Journal of Oceanic Engineering*, vol. 31, no. 4, pp. 904-918, 2006.
- [2] C. G. Quentin, Y. Barbin, L. Bellomo, P. Forget, D. Mallarino, J. Marmain, A. Molcard and B. Zakardjian, "High Frequency Surface Wave Radar in the French Mediterranean Sea: an element of the Mediterranean Ocean Observing System for the Environment," in *7th EuroGOOS Conference*, Lisboa, Portugal, 2014.
- [3] D. Golubović, M. Erić, N. Vukmirović and V. Orlić, "High-Resolution Sea Surface Target Detection Using Bi-Frequency High-Frequency Surface Wave Radar," *Remote Sensing*, vol. 16, no. 18, p. 3476, 2024.
- [4] D. Golubović, "The Future of Maritime Target Detection Using HFSWRs: High-Resolution Approach," in *2024 32nd Telecommunications Forum (TELFOR)*, Belgrade, Serbia, 2024.
- [5] D. S. Ilcev, "Introduction to coastal HF maritime surveillance radars," *Polish Maritime Research*, vol. 26, no. 3, pp. 153-162, 2019.
- [6] S. J. Anderson, "Optimizing HF Radar Siting for Surveillance and Remote Sensing in the Strait of Malacca," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 3, pp. 1805-1816, 2013.
- [7] D. Golubović, N. Vukmirović, M. Erić and M. Simić-Pejović, "Method for Noise Subspace Determination in HFSWR's High-Resolution Range-Doppler Map Estimation," in *2023 10th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN)*, East Sarajevo, Bosnia and Herzegovina, 2023.



- [8] D. Golubović, M. Erić and N. Vukmirović, “Improved Detection of Targets on the High-Resolution Range-Doppler Map in HFSWRs,” in *2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia and Herzegovina, 2024.
- [9] D. Golubović, N. Vukmirović and M. Erić, “An Introduction to Vessel Tracking in HFSWRs Based on a High-Resolution Range-Doppler Map: Some Preliminary Results and Challenges,” in *2024 13th Mediterranean Conference on Embedded Computing (MECO)*, Budva, Montenegro, 11-14 June 2024.
- [10] L. Sevgi, “Modeling and simulation strategies in high frequency surface wave radars,” *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 18, no. 3, pp. 399-408, 2020.
- [11] H. Roarty, T. Cook, L. Hazard, D. George, J. Harlan, S. Cosoli, L. Wyatt, E. Alvarez Fanjul, E. Terrill, M. Otero, J. Largier and e. al, “The Global High Frequency Radar Network,” *Frontiers in Marine Science*, vol. 6, p. 164, 2019.
- [12] C. A. Guérin, M. Rozel, D. Dumas, M. Menelle, V. Morales-Marquez and L. Perus, “<https://hal.science/hal-04653841v1>,” 19 7 2024. [Online]. Available: <https://hal.science/hal-04653841v1/document>. [Accessed 20 2 2025].
- [13] A. L. López, G. F. Segovia, R. V. Valencia and E. Q. Valencia, “Study of high frequency surface wave radar (HFSWR),” *Revista Ciências Exatas*, vol. 11, no. 2, 2005.
- [14] S. J. Anderson, “HF radar network design for remote sensing of the South China Sea,” in *Advanced Geoscience Remote Sensing*, IntechOpen, 2014.
- [15] D. Golubović and D. Marjanović, “Ship Detection Software Based on High-Resolution Algorithms for Primary Signal Processing: From Idea to Implementation in HFSWRs,” in *2025 24th International Symposium INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia and Herzegovina, 2025.
- [16] D. Golubović, N. Vukmirović, Z. Lončarević, M. Marković and M. Erić, “Execution Time Improvement using CPU Parallelization and Non-Uniform High-Resolution Range-Doppler Map Estimation in HFSWR,” in *9th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN 2022)*, Novi Pazar, Serbia, 6-9. June 2022.



CULTURAL TOURISM IN THE AGE OF EMERGING TECHNOLOGIES

Sava Cipriana*,
[0000-0002-6521-6596]

Dragotă Tiberiu–Gabriel
[0009-0009-1401-9240]

Tibiscus University Timișoara,
Timișoara, Timiș

Abstract:

Tourism is a practice that adds to culture, preserves the natural world, and offers peace. It is a human desire to make use of the spare time away from one's home. Regardless of the main reason for travelling, cultural interest related to the destination prevails.

People have always been fascinated by their own past such as different rituals that have contributed substantially to the evolution of civilisation. However, traditional modalities of discovery – books, museums, or documentaries – do not offer any experiential possibilities. It is in such a scenario that Augmented Reality (AR) and Virtual Reality (VR) enter the limelight. These two inventions have the potential to make history come alive, by actually transporting people into the worlds of the past that have been hidden for centuries, or to enable people to engage directly with their cultural heritage and traditions.

A VR or AR app could bring history to life – imagine walking through a mythical world, or in a medieval town where ancient skills can be learnt. This is the reason why these technologies are more than games; they are becoming fantastic learning tools and culture preservers.

Keywords:

Cultural Tourism, Technology, Augmented Reality, Virtual Reality, Unity, Unreal Engine.

INTRODUCTION

There is an increased use of immersive technologies in cultural tourism in current literature. An extensive review of the field by summarizes the AR and VR tourism trends deliver emotional and learning experiences to tourists [1]. Similarly, [2] comparing AR and VR's impact on engagement in the case of heritage sites emphasizes that AR forms contextual enrichment and VR as immersion of depth.

Extended Reality (XR) technology, including Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR), is increasingly gaining momentum in tourism and education. A review depicts XR's educational significance in cultural heritage education in 2024 [3], and a study discusses XR's capacity to involve tourists in intangible cultural heritage [4].

Correspondence:

Sava Cipriana

e-mail:

cipriana.sava@gmail.com





Recent trends in cultural tourism have seen the integration of innovative technologies beyond the traditional Extended Reality (XR) applications. Artificial Intelligence (AI) plays a pivotal role in enhancing tourist experiences through personalized recommendations and conversational agents, improving cultural interactions and making them more personalized. The Internet of Things (IoT) enables the development of intelligent tourism infrastructures, ensuring real-time collection and analysis of data to enrich visitor management and safety at cultural heritage sites [5]. Geolocation-enabled mobile apps with gamification and Augmented Reality (AR) features have become the norm in interactive storytelling, enabling tourists to experience cultural tales interactively [6]. Besides, blockchain technology is also researched to authenticate and trace digital cultural assets' provenance, ensuring a secure platform to handle and share cultural heritage information [7].

Besides technological innovation, digitization efforts also involve digitizing records, creating interactive software, and devising digital storytelling approaches that recontextualize heritage narratives for international consumption. Platforms such as Google Arts & Culture and Europeana have made museum collections and heritage sites available for everyone to access, which brings more visibility to marginalized cultures [8]. Social media platforms support participatory heritage-making, where users can share and co-author cultural narratives, which in turn creates a shared memory and involvement with cultural heritage [9]. The application of digital twins—digital replicas of actual cultural sites—has been of especial value in conservation activities and virtual tourism, especially when the travel is physically restricted, such as during the COVID-19 pandemic [10]. Combined, these digital tools not only save endangered cultural heritage but also enhance the tourist experience by making it more interactive, individualized, and accessible.

Panoramic VR content gamified storytelling has been put forth as a strong means in experiential tourism for engaged cultural immersion [11]. Simultaneously, the economic value of immersive tourism grows in visibility. Market research in 2024 anticipates a compound annual growth rate (CAGR) of 38% of VR tourist market through 2028 [12].

These findings warrant the strategic use of immersive technologies for cultural tourism as an educational technology and a tool for sustainable regional growth.

Tourism has changed over decades and centuries due to a variety of external factors – the most important being the evolution of transport and technology. Shorter travel times have resulted in longer durations of stay at destinations and greater engagement with local cultures.

In this context, tourism is not only a leisure activity but also an important motivation for travel, serving to meet spiritual needs and support cultural development. It is based on anthropic resources – including historical, religious, artistic, literary, traditional, architectural or culinary elements – and is expressed in a wide variety of forms.

Cultural tourism continues to attract an increasing number of travellers worldwide — both those who travel specifically for cultural purposes and those who engage with culture more incidentally. A major trend in this sector is the growth of unique experiences and meaningful interactions with local cultures.

Nowadays, we can assert that there is heated competition among tourism destinations, and technology may be a decisive factor that determines tourists' touring activities.

Augmented Reality, or AR for short, refers to an interactive experience of superimposing digital content on the real world. One can experience this type of reality with a smartphone, tablet, or even smart glasses. These devices use their cameras and sensors to recognize the physical environment and add virtual elements in real time. According to Azuma, AR systems must combine real and virtual environments, be interactive in real time, and register content in 3D space [13].

One sector with significant potential that has already begun adopting such solutions is cultural tourism. By scanning points of interest, visitors can discover the history of a location and how it used to look in the past. They may also encounter mythological characters, animations, or interactive guides. A clear example is illustrated in Figure 1 and Figure 2: tourists visiting Casa Batlló in Barcelona have the option to take a tour using tablets equipped with augmented reality features. Paintings, furniture, and many other elements within the museum come to life, revealing stories that cannot be interpreted by the human eye alone.

Traditional villages, with their rich historical heritage, can adopt this method to share their cultural values with visitors. Traditional folk dances, combined with myths and stories, can help create meaningful connections between tourists and the place. As Giaccardi argues, digital heritage experiences such as these enable users to actively participate in the meaning-making process of cultural sites [14]. An example is shown in Figure 3, which demonstrates how Augmented Reality (AR) can be used to digitally reconstruct ancient ruins for archaeological and cultural interpretation [15].

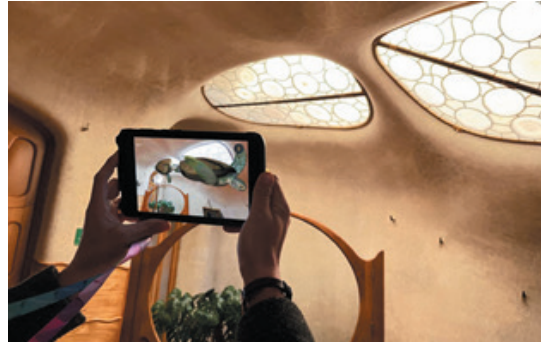


Figure 1. Viewing historic interiors through AR at Casa Batlló



Figure 2. Viewing historic interiors through AR at Casa Batlló



Figure 3. Examples of AR applications for archaeological purposes

According to Mize.tech [16], AR is increasingly used in tourism to deliver layered digital storytelling and enhance real-time visitor experiences. Using Augmented Reality in tourism is a promising approach since most smartphones and tablets support AR and it is relatively easy to integrate with the real world. The drawbacks concern the restrictions of immersion and reliance on the physical context of surroundings.

Virtual Reality (VR), on the other hand, is an immersive technology which places the user in a fully digital environment, separated from the real world. This is made possible through specialised VR helmets, which provide a 360-degree visual experience and allow full interaction with virtual environments via controllers or motion sensors. Slater and Wilbur emphasise the importance of “presence” in VR – the feeling of being physical-



ly situated within a simulated world [17]. An example of this can be seen in Figure 4, where a soldier uses virtual reality to simulate parachute jumps in a controlled and immersive environment.

In the context of cultural tourism, Virtual Reality (VR) enables the reconstruction of entire locations or historical places. Users can take part in traditional ceremonies and explore historical locations that no longer exist in the present days. These types of applications can be used in museums, in education, or even for personal use.

VR's strengths lie in its high degree of immersion and its virtually limitless creative potential. However, challenges remain, such as the high cost of equipment and the need for technical optimisation to ensure smooth performance.

The broader category of Extended Reality (XR) encompasses both Augmented Reality (AR) and Virtual Reality (VR), which operate in distinct ways to fulfil different purposes and contexts. Augmented Reality is well-suited for integration in real-world settings to deliver contextual information which enhances physical reality. On the other hand, Virtual Reality functions as a self-contained virtual universe that provides optimal conditions for complex simulations and immersive storytelling.

The proposed case study illustrates the differences between these two technologies, their development and preparation process, and their application in cultural tourism – using the theme “Magic Meadow”. Through AR, tourists are guided through real forests where mythical creatures appear. In VR, users can experience a digital version of the legendary meadow from their actual location.

2. CASE STUDY: THE “MAGIC MEADOW” EXPERIENCE AS A DIGITAL HERITAGE MODEL

All countries preserve elements of their folklore and mythology, even if only in fragments.

Throughout Romania, various legends are passed down, including stories of “Sânziene”, “lele”, fairies, “Baba Dochia”, “Ileana Cosânzeana” and beautiful maidens who have turned into waterfalls. Other mythical figures include “Iovan Iorgovan”, werewolves, and more. There are also wonderful natural sites, as well as man-made sites – manors, fortresses, and ancient ruins – where one can still feel the echo of tales whispered across centuries.

Some of these places continue to attract tourists, such as Poiana Narciselor (Daffodil Meadow) in Zervești, Caraș-Severin County, a nationally protected area settled in IUCN Category IV, that is chosen as a botanical nature reserve. The reserve spans approximately 40 hectares and lies within the administrative territory of Turnu Ruieni Commune. The road distance from Timișoara to Zernești is about 116 km, via Timișoara – Caransebeș–Turnu Ruieni route. From Caransebeș to the village of Zernești the distance is only 6 km, with access to the protected area made via an agricultural road.

The Festival of the Daffodils was first held here in 1965 and it usually takes place in May, when the white flowers are in full bloom. Each year, the local authorities organize open-air celebrations to promote regional traditions, including local attire, traditional dance, and specific culinary heritage.



Figure 4. Soldier using VR for parachute jumps simulation



According to local folklore, the Roman poet Ovid is said to have passed through this area on his way to exile. Upon seeing this stunning field of narcissus, he is believed to have exclaimed: “Cara mihi sedes” — “My beloved home.”

It is a peaceful glade, bathed in sunlight by day and moonlight by night – a place where one feels as though they are stepping into a fairytale.

The application “Magic Meadow”, inspired by Daffodil Meadow in Zervești, Caraș-Severin Country, serves as an example of how storytelling can be used to engage tourists with the history of a place, incorporating mythological content and traditional practices. Tourists are transported into a mythological world, where they may encounter fairies as well as legendary stars and talking trees. Every part of this fictional realm is rooted in stories drawn from Romanian folklore, alongside traditional customs and historical spiritual practices. According to Green and Brock (2000), narrative transportation – the feeling of being absorbed in a story – enhances the impact of such experience [18].

The creation of such an application requires multiple development phases. Collaboration with ethnographers and local representatives is essential during the documentation stage. A narrative map of the meadow will be created, identifying key points where stories will be triggered.

The following development stages require the construction of a 3D environment using Unity or Unreal Engine as graphic engines. The environment requires 3D models of trees and flowers and mythical creatures together with special effects including fog and volumetric lighting and magical particles to establish a mythical atmosphere. Users will be able to interact with elements through proximity detection, which will activate audio stories and animations.

The audio component plays a crucial role in enhancing the immersive experience. Spatial sound techniques can be used to simulate natural and magical elements — such as birdsong, wind, or whispers — coming from different directions. Storytellers can narrate the tales in a theatrical style, accompanied by music inspired by local traditions.

The application can be developed in both an AR version, accessible via smartphones or tablet, and a VR version, designed for a fully immersive experience using headsets.

3. TECHNICAL ARCHITECTURE: PLATFORMS, LANGUAGES, AND DEVELOPMENT TOOLS

The development of immersive cultural tourism applications relies on advanced engines and development frameworks capable of delivering interactive, high-fidelity, and emotionally resonant experiences. The two primary engines in use—Unity and Unreal Engine—offer robust toolsets for building both AR and VR environments and are widely adopted across educational and heritage-focused XR projects.

Unity is favored for mobile AR applications due to its flexibility, large developer community, and support for cross-platform deployment. It integrates seamlessly with AR Foundation, a unified framework supporting both ARKit (iOS) and ARCore (Android), as well as Vuforia for image-based AR experiences. Developers leverage Unity’s C# scripting, NavMesh Agents for character movement, Collider systems, i.e. Listing 1, and XR Interaction Toolkit for real-time interactions. Recent research highlights Unity’s effectiveness in rapid prototyping of educational XR apps and integrating multimodal user interaction—such as gesture, voice, and proximity-triggered storytelling.

Unreal Engine, on the other hand, provides superior graphical fidelity, making it ideal for large-scale VR environments and photorealistic cultural simulations. Its Blueprints visual scripting system supports fast iteration without deep coding expertise, while the Niagara VFX System enables dynamic particle effects like glowing creatures, fireflies, and magical trails. The Sequencer tool allows complex storytelling scenes to unfold in synchrony with sound, animation, and environmental changes. Unreal Engine is also recognized for its support of MetaHuman characters, making it suitable for simulating virtual docents or mythological figures in high-detail.

```
void OnTriggerEnter(Collider other) {
    if (other.CompareTag("Player")) {
        fairyEffect.Play();
        narratorAudio.Play();
    }
}
```

Listing 1. An example of a OnTriggerEnter() function written in C#



Further enhancements include the use of spatial audio, haptic feedback systems (e.g., for interactive museum exhibits), and AI-driven NPCs using Unity ML-Agents or Unreal's Behavior Trees to simulate interactive folklore characters with autonomous reactions. The OpenXR standard ensures compatibility with a broad range of headsets.

Developers are also incorporating cloud rendering (e.g., NVIDIA CloudXR) to offload heavy visual computation and make rich XR experiences accessible on lightweight mobile devices. Moreover, spatial computing platforms are shaping a new generation of cultural applications rooted in environmental understanding and gesture-based input.

Together, these technologies offer a powerful toolset for building highly immersive, culturally respectful, and technically optimized tourism experiences that are deployable across various devices and environments.

3.1. INTERACTION DESIGN: USER EXPERIENCE IN AR AND VR ENVIRONMENTS

The proposed application functions as an extension of real-world locations – such as forests and ethnographic gardens – within the AR version. Users can experience actual environments through their mobile phones or tablets, while the screen displays interactive virtual content: fairies flying through trees, legendary stags running across clearing, ancient symbols floating in the air, and magical story gateways. The system responds to user input through physical object proximity detection, as well as screen touch and voice commands. The physical path guidance users augmented interactive maps which provide direction while voice-over narration and 3D sound effects deliver the stories. The system targets authentic location visitors who want to enhance their experience without compromising their connection to nature or local heritage.

The VR version provides a fully digital recreation of the entire experience which operates independently from the user's location. The virtual environment includes detailed simulations of natural elements – such as trees, plants, and animals – together with magical atmospheric effects such as auroral lights and mythical mist. Users experience a dreamlike meadow through VR headsets while they explore freely and interact with magical characters, trigger narrative events, and collect symbolic objects. VR also facilitates learning through interaction, aligning with constructivist educational theories [19].

The virtual world can include secret places, gateways to another dimension, and seasonal changes that evolve over time. In addition, virtual reality allows for the incorporation of interactive gaming elements – including tasks to complete, mysteries to unravel, or recreations of ancient rituals. This variant is appropriate for individual exploration and learning exhibition settings, here users can engage with a highly creative narrative that meets cultural expectations comprehensively.

3.2. DEPLOYMENT AND PERFORMANCE OPTIMIZATION

The final stage involves enhancing the performance of the application through optimisation. Techniques such as Level of Detail (LOD), light baking, and texture compression are used. Low-level testing guarantees the readiness of the application for publishing on distribution channels such as SideQuest and SteamVR for the virtual reality and Google Play and App Store for the augmented reality version.

3.3. ETHICAL AND CULTURAL CONSIDERATIONS IN DIGITAL HERITAGE

The application's design must remain sensitive to the cultural context it represents. The active involvement of local communities is necessary to maintain the authenticity of the content. As noted by UNESCO, safeguarding intangible cultural heritage requires ethical representation and community involvement [20].

The use of stereotypes should be avoided, and the commercialisation of heritage should be restricted. In addition, the application should be made accessible and inclusive, with multilingual options and facilities for disabled users.

3.4. BROADER IMPLICATIONS: AR/VR IN CULTURAL TOURISM AND EDUCATION

Implementing an application of this kind – whether it is AR or VR – within the cultural tourism sector is more than a simple recreation of an immersive experience. Such uses are a change in basic assumptions in the way we learn, engage with history, and experience cultural heritage. At a macro level, they can also serve as strategic instruments of regional development, revitalisation of local identity, and activation of creative economies.



From a learning perspective, extended reality enables learning through experience; children get to learn about myths and traditions not just through texts but by being part of those experiences. XE becomes a strategic tool for regional identity, creative economies, and sustainable tourism [14].

Technologically, the rapid evolution of graphics processing, artificial intelligence, and 5G will enable ever more realistic, interactive, and immersive experiences. According to Statista, the global extended reality market will surpass USD 100 billion by 2026 [21], a large increase compared to 2023 when it was USD 41.2 billion.

Culturally, digitisation of myths and rituals by interactive media can prevent their disappearance, but it must also be ethically responsible. There needs to be an assurance that digital representations do not simplify or distort original meanings. This calls for collaboration among developers, artists, ethnographers, and local communities.

In tourism, these kinds of applications like the one proposed can help increase the attractiveness in rural areas, offering sustainable options to mass tourism and being attractive to customers seeking different and authentic experiences. Moreover, UNWTO data from the World Tourism Barometer indicates that cultural and interactive experiences remain a priority for a growing number of international travellers [22].

4. CONCLUSIONS: EXTENDED REALITY AND THE FUTURE OF CULTURAL HERITAGE ENGAGEMENT

The combination of AR/VR with tourism is not just a technological development – it is a revolutionary new way of preserving, understanding, and experiencing heritage. With virtual reality rooted in local myth and legend, such apps call for a unique coming together of past and present, virtual, and real. Whether through AR, which is augmenting existing environments, or VR which is immersing users in completely new ones, extended reality is yielding enormous tools for education, preservation of cultures, and establishing sustainable tourism.

Developers should take into consideration cultural sensitivity, attentiveness to detail, and ethical responsibility. Respect for traditional knowledge, participatory communal engagement, and authenticity are imperatives in bringing digital heritage from the shallow reproduction making it a profound extension of cultural memory.

As technology keeps evolving, it also enhances the potential to shape the future of cultural engagement – a culturally engaged interaction that is interactive, inclusive, and in close resonance with the values of the cultures that it seeks to represent.

REFERENCES

- [1] M. A. Bretos, S. Ibáñez-Sánchez and C. Orús, "Applying Virtual Reality and Augmented Reality to the Tourism Experience: A Comparative Literature Review," *Spanish Journal of Marketing - ESIC*, vol. 27, no. 2, pp. 145-165, 2023.
- [2] I. Pinto and A. Huertas, "A comparative study of VR and AR heritage applications on visitor emotional experiences: a case study from a peripheral Spanish destination," *Virtual Reality*, vol. 29, p. article no. 36, 2025.
- [3] A. Dordio, E. Lancho, M. Merchán and P. Merchán, "Cultural Heritage as a Didactic Resource through Extended Reality: A Systematic Review of the Literature," *Multimodal Technologies and Interaction*, vol. 8, no. 7, p. 58, 2024.
- [4] M. S. Anwar, J. Yang and J. Frnda, "Metaverse and XR for cultural heritage education: applications, standards, architecture, and technological insights for enhanced immersive experience.," *Virtual Reality*, vol. 29, p. article no. 51, 2025.
- [5] Ü. Gaberli, "Cultural Tourism Internet of Things and Smart Technologies in Museums," *Handbook of Research on Digital Communications, Internet of Things, and the Future of Cultural Tourism*, pp. 260-270, 2022.
- [6] E. Ceh-Varela and G. A. Hernandez-Chan, "Review of Mobile Applications for Tourism," *International Journal of Electronics and Computer Science Engineering IJECSE*, pp. 353-356, 2015.
- [7] H. Stublić, M. Bilogrivić and G. Zlodi, "Blockchain and NFTs in the cultural heritage domain: A review," p. 3801-3819, 2023.
- [8] N. P., "Museums and the Post-Digital: Revisiting Challenges in the Digital Transformation of Museums.," pp. 1784-1800, 2024.
- [9] M. Economou, "Heritage in the digital age. A Companion to Heritage Studies," pp. 215-228, 2015.
- [10] J. Hutson, J. Weber and Russo, "A. Digital Twins and Cultural Heritage Preservation: A Case Study of Best Practices and Reproducibility in Chiesa dei SS Apostoli e Biagio.," p. 15-41, 2023.
- [11] E. Kontogiorgakis, E. Zidianakis, E. Kontaki, N. Partarakis, C. Manoli, S. Ntoa and C. Stephanidis, "Gamified VR Storytelling for Cultural Tourism Using 3D Reconstructions, Virtual Humans, and 360° Videos," *Technologies*, vol. 12, p. 73, 2024.



- [12] GlobeNewswire, "Virtual Reality in Tourism Market Report 2024," GlobeNewswire, pp. <https://www.globenewswire.com/news-release/2024/11/13/2980386/28124/en/Virtual-Reality-in-Tourism-Market-Report-2024.html>, 13 November 2024.
- [13] R. T. Azuma, "A survey of augmented reality.," *Teleoperators & Virtual Environments*, vol. 6, no. 4, pp. 355-385, 1997.
- [14] E. Giaccardi, *Heritage and Social Media: Understanding Heritage in a Participatory Culture*, Routledge, 2012.
- [15] B. M., R. Pierdicca, F. E. and M. E. S., "A Survey of Augmented, Virtual, and Mixed Reality for Cultural Heritage," *Journal on Computing and Cultural Heritage*, pp. 1-36, 2018.
- [16] Mize.tech, "5 Ways Augmented Reality is Enhancing the Tourism Experience," Mize.tech, pp. <https://mize.tech/blog/5-ways-augmented-reality-is-enhancing-the-tourism-experience/>, 01 04 2025.
- [17] M. Slater and S. Wilbur, "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence," *Presence: Teleoperators & Virtual*, vol. 6, no. 6, pp. 603-616, 1997.
- [18] M. C. Green and T. C. Brock, "The Role of Transportation in the Persuasiveness of Public Narratives," *Journal of Personality and Social*, vol. 79, no. 5, pp. 701-721, 2000.
- [19] S. Papert, *Children, Computers, and Powerful Ideas*, Basic Books, 1980.
- [20] UNESCO, "Convention for the Safeguarding of the Intangible Cultural Heritage," 2003. [Online]. Available: <https://ich.unesco.org/en/home>.
- [21] Statista, "Extended Reality (XR) Market Size Worldwide 2020–2026," 2023. [Online]. Available: <https://www.statista.com/topics/11185/mixed-reality/>.
- [22] UNWTO, "UN Tourism World Tourism Barometer – Latest Data," 2023. [Online]. Available: <https://www.unwto.org/un-tourism-world-tourism-barometer-data>.



LEARNING A FIRST PROGRAMMING LANGUAGE WITH LLM TUTOR

Milić Vukojičić *,
[0009-0002-1218-5893]

Ivana Korica
[0009-0009-1918-347X]

British International School,
Belgrade, Serbia

Abstract:

The paper presents the use of Large Language Models (LLM) and Chatbots as tutors when students need to learn novel concepts related to programming. Traditional methods of teaching and learning new programming concepts involve the students and teachers exploring one idea at a time – sequence, selection, and iteration – followed by more complex concepts. Traditional methods include using textbooks, computers, presentations, etc., and the teacher has the central role in the classroom. LLMs can help make complex topics more accessible for teachers to teach and for students to explore. Passive lectures and frontal teaching methods are replaced by the interactive use of LLM/chatbots, where the teacher has a central role in checking the information produced by the LLM/chatbot and is the key mediator between students and the LLM. The study was conducted with 30 students (n=30) in two groups. One group used LLM/chatbot as the core resource for first-time learning the programming language (n=15), while the second group used standard teaching and learning methods (n=15). The results show a significant improvement for first-time learning the programming language with the help of LLM – students showed better results in the assessment, besides being happier with the experience and the learning process. The study also demonstrates potential problems with this method, such as time, resources, accessibility issues, and initial preparation.

Keywords:

Large Language Models, Chatbots, Secondary Education, Programming Languages, Artificial Intelligence.

INTRODUCTION

The integration of Artificial Intelligence (AI) into computer science education presents a transformative potential for teachers and students alike. Modern curricula that are based on critical thinking and problem-solving can benefit from tools like Large Language Models (LLM) and chatbots as an opportunity for students to learn critical thinking and problem-solving and for teachers to implement these fields in their subjects easily. Modern lower secondary education often struggles to equip students with fundamental programming concepts such as programming syntax, control structures (e.g. loops, conditional statements), and the skill to debug their code. This is mainly due to the complexity of the tasks and the limited time available for more serious practice in the lessons.

Correspondence:

Milić Vukojičić

e-mail:

vukojicic.milic@gmail.com





The approach of using LLM to help students generate code can be seen in various papers. Some papers [1] agree that standard methods of learning how to code can be demanding because of the way that instructions are presented, often with robust textual explanations. Studies also suggest that tools like LLM can help us with the power of the Integrated Development Environment and make a significant impact when it comes to writing computer programs.

Other studies suggest [2] that LLM can be used as a programming assistant by helping students write better and more effective code. This means that students can write code with a better understanding of what they are writing, and it also suggests that students appreciate other parts of code as very valuable, for example, comments. We can also see that in the comprehensive study, there are many benefits of using chatbots in education [3], as the study highlights the benefits of using chatbots as a virtual assistant.

The impact of LLMs within the domain of undergraduate education, secondary and lower secondary education is still limited. Some of the results are not in favour of LLMs in the domain of higher education [4] in addition to the domain of lower secondary education [5]. We can see a negative correlation between using LLMs and lower grades, in addition to lower use of critical thinking in undergraduate education. Similarly, when referring to resources, we can also see a negative correlation between students' grades and the use of LLMs in lower secondary education. From the work mentioned above, we can see several potential advantages and disadvantages in regard to the extensive use of LLMs in the education system. This is why we believe that teachers, tutors, and professors must play a crucial role in the education system – they need to monitor the output of LLM software and guide students towards their proper use. (Figure 1).

The research suggests that educational institutions must promote the use of LLM models [6] in addition to encouraging the best practices. Chatbots and LLMs can be used in order to promote the exploration of novel ideas and new content, along with problem-solving skills and critical thinking. With the rapid movement to the area of LLMs, schools and educators should reconsider resizing and redefining their assessment standards and student evaluation standards.

This experiment will introduce the idea of using LLMs for learning a programming language for the first time, as students will explore concepts like sequence, selection, and iteration. All of the students chose Python as the first programming language, and LLMs will be used as tutors, while the teacher will fill the role of a mediator between LLMs and students.

2. EXPERIMENT METHODOLOGY

This experiment was based on observing and analysing the work of 30 lower secondary students, aged between 10 and 14, of different genders and a variety of backgrounds, with zero experience in textual programming languages. The students were organised into two groups. The first group used LLM/chatbot as a tutor, and the second group used standard methods of education such as textbooks, notebooks, slides, and teacher presentations. This paper will compare the results of both groups on the assessment. The assessment was constructed of 9 questions, with 3 questions from each category, starting from low-level question (1 point), medium-level question (2 points) to high-level question (3 points):

- writing sequences with Python programming language;
- using Python programming language to define selection (if-statement);
- defining iterations in Python programming language (for and while loops).

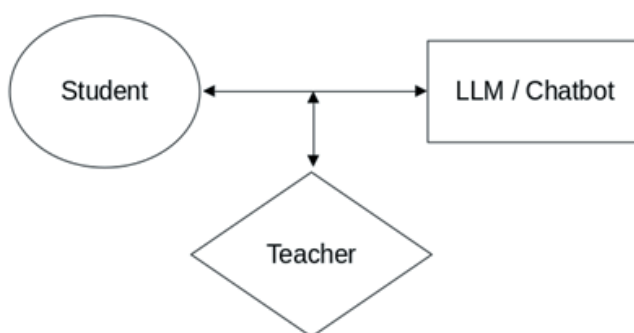


Figure 1. LLM, student and teacher relationship



Group 1 had the training on how to use LLM/chatbot in the domain of computing education. Students could choose between several LLM models: ChatGPT 4o-mini, Llama 3.3 70B, Claude 3 Haiku, and Mistral Small 3. The students also used the recommended prompting method: Persona, Aim, Recipients, Theme, and Structure, as suggested by Google and other researchers [7]. The different areas of prompt are defined as:

- Persona – Students will define who they are;
- Aim – Students will define a specific task for the prompt to solve;
- Recipients – they will define themselves because they are the recipient of the task;
- Theme – this is where LLM needs to set the voice, for example, to be more formal than usual;
- Structure – instruct the AI that it needs to behave as a tutor, with what it needs to explain and how it will ask questions about certain topics.

An example of a student's prompt: "I'm a Year 7 Computing student. Pretend that you are the Computer Science teacher. You want me to learn the Python programming language, you need to lead me to the key concepts like sequence, selection, and iteration, starting from the basic knowledge of Python language. After your explanation, you will need to ask me a question about the topic that you have explained. You will need to evaluate my answers. Each interaction should have an explanation of one concept and one question. After evaluation, you will teach me the next concept.". In Figure 2, we can see an example of the output of the LLM and the interaction with the user.

3. Results and discussion

Results are obtained from all 30 students after completing all 9 questions, 3 questions from each category, starting from low-level question (1 point), medium-level question (2 points) to high-level question (3 points) about the topics mentioned before. The data is shown in Table 1, with the results of Group 1 and Group 2.

In Table 2, we can see the results of both groups of students in the different question levels. We can see that group 1, which used an LLM/chatbot, has better results in questions about sequence and selection but lower results from iteration from low-level questions. In the medium-level question category, Group 2 had better results from questions related to sequence and lower performance with questions related to selection and iteration. In the domain of high-level questions, Group 1 achieved higher scores.

In Table 3, we can observe the lowest test score expressed in percentages from each group, which is 22% for Group 1 and 22.22% for Group 2. The highest score achieved in the test is 100% in both groups. The average score on the test for Group 1 is 71,48%, and for Group 2 is 65,18%.

From Table 1, Table 2, and Table 3, we can see that Group 1, the group that used an LLM/chatbot as a tutor, had better results on the test than Group 2, which used traditional methods of learning in the process of learning a textual programming language for the first time.

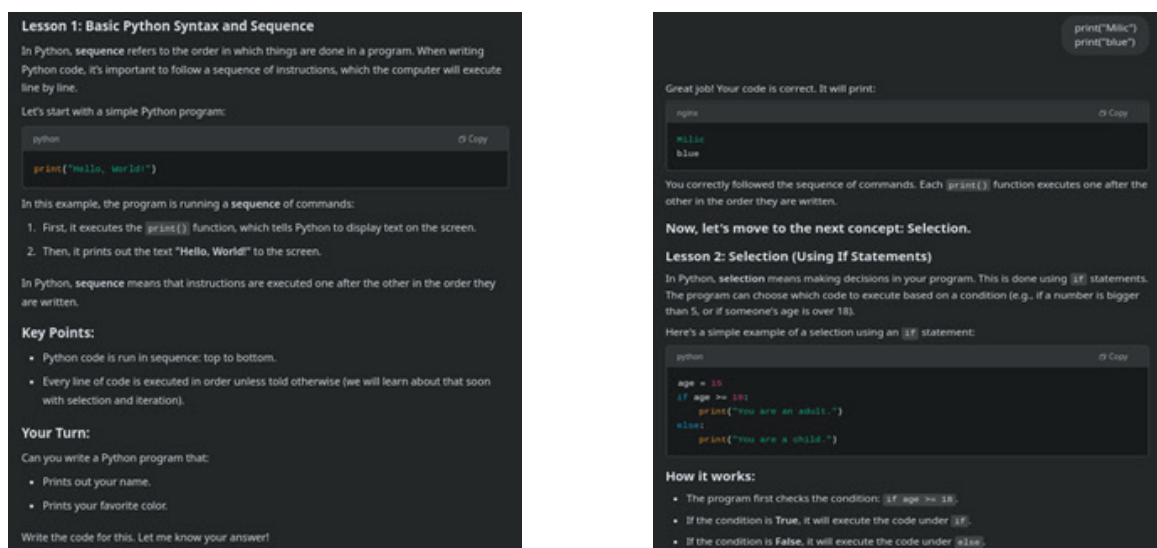


Figure 2. LLM output and the example of an interaction with the user

**Table 1.** Results of Students (S1- S15) separated into Group 1 -LLM (G1) and Group 2 – without LLM (G2), with different levels of question difficulty (low, medium, high) per each category (sequence, selection, iteration)

	Low -level question			Medium-level question			High-level question		
	Sequence	Selection	Iteration	Sequence	Selection	Iteration	Sequence	Selection	Iteration
S1 (G1)	0	0	0	1	1	1	1	0	0
S2 (G1)	1	1	1	2	2	2	3	3	3
S3 (G1)	1	1	1	2	2	2	3	3	3
S4 (G1)	1	1	0	2	1	1	3	0	1
S5 (G1)	1	1	1	2	0	1	3	2	2
S6 (G1)	1	0	0	2	1	1	3	2	3
S7 (G1)	1	0	0	2	0	0	3	1	2
S8 (G1)	1	1	1	2	2	1	3	3	1
S9 (G1)	1	1	0	1	2	2	2	1	1
S10 (G1)	1	0	1	2	0	0	3	1	1
S11 (G1)	1	1	1	2	2	2	3	3	3
S12 (G1)	1	1	0	2	2	1	3	1	1
S13 (G1)	1	1	0	2	2	2	3	2	2
S14 (G1)	1	1	1	2	2	2	2	3	2
S15 (G1)	1	0	1	2	1	1	2	2	2
S1 (G2)	1	0	0	2	1	1	3	1	0
S2 (G2)	1	1	0	2	1	0	2	1	1
S3 (G2)	0	1	1	2	2	2	3	2	3
S4 (G2)	1	0	1	2	1	2	3	3	3
S5 (G2)	1	1	1	2	2	1	3	2	1
S6 (G2)	1	1	1	2	2	2	3	3	3
S7 (G2)	1	0	1	2	0	0	3	1	1
S8 (G2)	1	1	1	2	1	0	3	2	0
S9 (G2)	1	0	0	2	0	0	1	0	0
S10 (G2)	1	0	0	2	1	1	2	2	2
S11 (G2)	1	0	1	2	2	2	2	1	2
S12 (G2)	1	0	0	2	1	2	1	1	2
S13 (G2)	1	1	1	2	2	1	3	2	1
S14 (G2)	1	1	1	2	1	1	2	2	2
S15 (G2)	0	1	0	2	1	0	3	1	1

Table 2. Comparison between results (Group 1 - G1, Group 2 – G2) of all questions

	Low -level question			Medium-level question			High-level question		
	Sequence	Selection	Iteration	Sequence	Selection	Iteration	Sequence	Selection	Iteration
Result (G1)	14	10	8	28	20	19	40	27	27
Result (G2)	13	8	9	30	18	15	37	24	22

Table 3. Test score criteria: Min, Max and Average (for all students) per group (Group 1 – G1, Group 2 -G2)

Test score criteria	%
Min (G1)	22
Min (G2)	22,22
Max (G1)	100
Max (G2)	100
Average (G1)	71,48
Average (G2)	65,18

**Table 4.** Overall satisfaction with learning (1-5) new concepts (sequence, selection, iteration)

Overall satisfaction with learning (1-5) new concepts			
	Sequence	Selection	Iteration
S1 (G1)	5	5	5
S2 (G1)	5	4	5
S3 (G1)	5	5	5
S4 (G1)	4	4	4
S5 (G1)	5	5	5
S6 (G1)	5	5	5
S7 (G1)	1	1	1
S8 (G1)	5	5	5
S9 (G1)	3	4	5
S10 (G1)	1	1	1
S11 (G1)	5	5	5
S12 (G1)	4	5	5
S13 (G1)	5	2	2
S14 (G1)	3	3	3
S15 (G1)	5	3	3
S1 (G2)	2	2	2
S2 (G2)	4	2	2
S3 (G2)	3	3	3
S4 (G2)	4	4	4
S5 (G2)	4	4	4
S6 (G2)	5	5	5
S7 (G2)	5	5	3
S8 (G2)	3	3	3
S9 (G2)	1	0	0
S10 (G2)	1	0	0
S11 (G2)	1	0	1
S12 (G2)	1	0	0
S13 (G2)	1	1	1
S14 (G2)	1	1	1
S15 (G2)	0	1	0

Table 5. Overall satisfaction with learning score

Overall satisfaction with learning score			
	Sequence	Selection	Iteration
Result (G1)	64	61	64
Result (G2)	56	51	50

From Table 4, we can see the overall satisfaction with learning, where each student rated their learning experience from 1 to 5 – 1 being “Not Satisfied at all” and 5 being “Completely Satisfied” related to the concepts of sequence, selection, and iteration. In Table 5, we can see the overall score for satisfaction, where the learning is better in group 1, which used LLM/chatbot as a tutor for learning new concepts.

4. CONCLUSION

In conclusion, this study provides fresh insights in the domain of using an LLM/chatbot as a tutor in the context of learning a new programming language. The study showed that there are significant differences between Group 1, which used LLM/chatbots as tutors and Group 2, which used traditional and ordinary methods of learning.



The average score of Group 1 is 71,48%, which is higher than the average score of Group 2, which is 65,18%. We can also see from the questionnaire that students find using an LLM/chatbot method more satisfying and that they had a better time and more fulfilment when we compared the results with those of Group 2.

The main problem that can occur is that student needs more time to adapt to a new method of learning and to the use of new tools in a meaningful way. Future work in the domain of LLMs and learning a programming language for the first time can be improved by looking at the methods for faster adaptation of students to their new learning environment and making students more independent in the domain of exploring new ideas, critical thinking, and problem-solving.

REFERENCES

- [1] D. Nam, A. Macvean, V. Hellendoorn, B. Vasilescu, and B. Myers, "Using an LLM to help with code understanding," in *Proc. IEEE/ACM 46th Int. Conf. Softw. Eng. (ICSE)*, Lisbon, Portugal, Apr. 2024, pp. 97:1–97:13. doi: 10.1145/3597503.3639187.
- [2] M. Vukojić and J. Krstić, "ChatGPT in programming education: ChatGPT as a programming assistant," *InspirED Teachers' Voice*, vol. 2023, no. 1, pp. 7–13, 2023. [Online]. Available: <https://www.teachers-voice.org/index.php/inspired/article/view/10>
- [3] C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," *Educ. Sci.*, vol. 13, no. 4, p. 410, Apr. 2023. doi: 10.3390/educsci13040410.
- [4] G. Jošt, V. Taneski, and S. Karakatić, "The impact of large language models on programming education and student learning outcomes," *Appl. Sci.*, vol. 14, no. 10, p. 4115, May 2024. doi: 10.3390/app14104115.
- [5] M. Vukojić, I. Korica, and M. Veinović, "The impact of LLM-based chatbots on secondary computing education," in *Sinteza 2024 - Int. Sci. Conf. Inf. Technol., Comput. Sci., Data Sci.*, Belgrade, Serbia: Singidunum Univ., 2024, pp. 31–37. doi: 10.15308/Sinteza-2024-31-37.
- [6] M. M. Rahman and Y. Watanobe, "ChatGPT for education and research: Opportunities, threats, and strategies," *Appl. Sci.*, vol. 13, no. 9, p. 5783, May 2023. doi: 10.3390/app13095783.
- [7] C. Chan, J. Cheng, X. Liu, Y. Yim, Y. Jiang, Z. Deng, H. Li, Y. Song, G. Y. Wong, and S. See, "Audience Persona Knowledge-Aligned Prompt Tuning Method for Online Debate," in *Proc. 27th Eur. Conf. Artif. Intell. (ECAI 2024)*, Santiago de Compostela, Spain, Oct. 2024, pp. 3851–3858. doi: 10.3233/FAIA240948.



MANAGEMENT AND TECHNOLOGY SESSION





ENHANCED MONTE CARLO SCHEDULE ANALYSIS: EVALUATION OF THE OPEN-SOURCE PERT-BASED SIMULATION TOOL

Goran Avlijaš*
[0000-0001-6379-3341]

Singidunum University,
Belgrade, Serbia

Abstract:

Monte Carlo simulations represent a powerful and beneficial method for risk analysis and decision-making, especially in project management. Traditionally, proprietary software such as Oracle Crystal Ball has been the go-to solution for Monte Carlo simulations, offering advanced features at a significant cost. Recently, web-based open-source tools, such as the PERT-based Monte Carlo simulation Tool, have emerged as a viable alternative. This study compares the performance of an open-source, web-based Monte Carlo simulation tool using PERT (Program Evaluation and Review Technique) methodology against Oracle Crystal Ball, a widely adopted commercial software. By analysing a simplified project schedule with varying levels of uncertainty, we evaluate the accuracy, usability, scalability, and cost-effectiveness of the open-source tool. Results demonstrate that the open-source PERT-based Monte Carlo simulation tool provides similar, if not superior, results in terms of accuracy and usability while offering significant advantages in terms of cost, accessibility, and adaptability.

Keywords:

Monte Carlo Simulation, Project Schedule, PERT, Open-source.

INTRODUCTION

Monte Carlo simulations are widely used for risk analysis and decision-making purposes in various fields, including finance, medicine, engineering, and project management [1], [2], [3], [4]. In the context of project management, Monte Carlo simulations help estimate the possible outcomes of project schedules based on the uncertainty of task durations [5]. Critical Path Method (CPM) and Program Evaluation and Review Technique (PERT) are often applied within Monte Carlo simulations to model the uncertainties in project task durations.

While commercial software like Oracle Crystal Ball has long been a leader in this area [6], it comes with a cost and limited access for smaller businesses or individuals. In contrast, recent advancements in open-source tools, specifically those based on web technologies, have provided an alternative solution that offers both cost efficiency and the ability to customize models. This paper aims to compare the performance of an open-source, web-based Monte Carlo simulation tool with Oracle Crystal Ball, focusing on factors such as accuracy, computational efficiency, user interface, and cost-effectiveness.

Correspondence:

Goran Avlijaš

e-mail:

gavlijas@singidunum.ac.rs





2. LITERATURE REVIEW

The use of Monte Carlo simulations in project management dates back to the 1950s, with key developments in modelling uncertainty and probabilistic risk analysis. Traditionally, tools like Oracle Crystal Ball have been regarded as industry standards for Monte Carlo simulations. The software integrates seamlessly with Microsoft Excel, offering sophisticated algorithms, predefined probability distributions, and robust graphical interfaces. Monte Carlo simulations have become an essential tool for assessing project schedule risks, and Oracle Crystal Ball is widely used for this purpose.

Despite its long-time existence, the Monte Carlo method is still insufficiently utilized in project management practices. Its use is primarily associated with cost and time management, where it helps quantify the level of risk associated with projected budgets or completion timelines. By simulating various scenarios, Monte Carlo analysis can reveal the probability of meeting planned deadlines or provide expected outcomes in terms of time and cost with a specific degree of reliability [7], [8]. This makes it a valuable tool for assessing uncertainties and making informed decisions.

In the context of time management, Monte Carlo simulation is particularly useful during the schedule development process, as it quantifies the confidence level of achieving targeted completion dates [9], [10]. Simulating project schedules using the Monte Carlo method is a cornerstone of quantitative risk analysis [11]. To obtain the most accurate activity duration estimates, industry experts assign probability distribution functions to each project activity.

A common approach is the use of three-point estimates (optimistic, most likely, and pessimistic), which are typically fitted to probability distributions such as normal, beta, or triangular distributions for each activity. This allows project managers to calculate the likelihood of meeting planned deadlines and establishment of schedule reserves on a data-driven basis. Monte Carlo simulations can be executed using dedicated project management software, such as Primavera Project Planner or Microsoft Project, or through simulation add-ins like Oracle Crystal Ball, which integrates with Excel for advanced probabilistic analysis.

Despite its powerful capabilities [12], Oracle Crystal Ball remains a costly solution with licensing fees that may not be feasible for all organizations and individuals. In response, several open-source Monte Carlo simulation tools have been developed in recent years. These

tools are generally accessible to a broader audience due to their low cost, open-source nature, and flexibility in modelling specific scenarios. PERT-based simulation tool developed by Christophe Yamahata represents one of these user-friendly alternative options that promise similar benefits due to its ability to model task uncertainties using three-point estimates [13].

3. METHODOLOGY

To evaluate the performance of the open-source Monte Carlo PERT-based simulation tool, a comparative analysis was conducted using data derived from a simplified project scenario. For the purpose of direct comparison with Crystal Ball, authors previous study [5] and input data was used to assess the tools' capability of handling the same project configuration. In order to avoid duplication results, this study will focus on illustration of results obtained by the open-source PERT-based tool.

The application of Monte Carlo simulation for project time management requires a strong understanding of key project management concepts, such as the Critical Path Method (CPM). CPM is a scheduling algorithm that calculates the early and late start and finish times for each project activity, identifies the critical path, determines activity floats, and computes other essential schedule parameters [14], [15].

In contrast to Oracle Crystal Ball, the open-source PERT-based tool enables the CPM diagramming technique depicted in Figure 1. Critical path in this example consists of activities A-D-J, and the shortest possible time to complete the entire project, given its predefined dependencies, is 43 days. This integration of Monte Carlo simulation with CPM enhances the ability to manage uncertainties and optimize project schedules effectively.

Both tools utilized beta-PERT distribution for task durations. Once the distributions are defined, random inputs are selected (typically thousands or millions of times) to generate a frequency distribution of the outcome variable. This outcome provides valuable insights, such as the probabilities of missing deadlines, completing the project ahead of schedule, and other critical information to support decision-making processes.

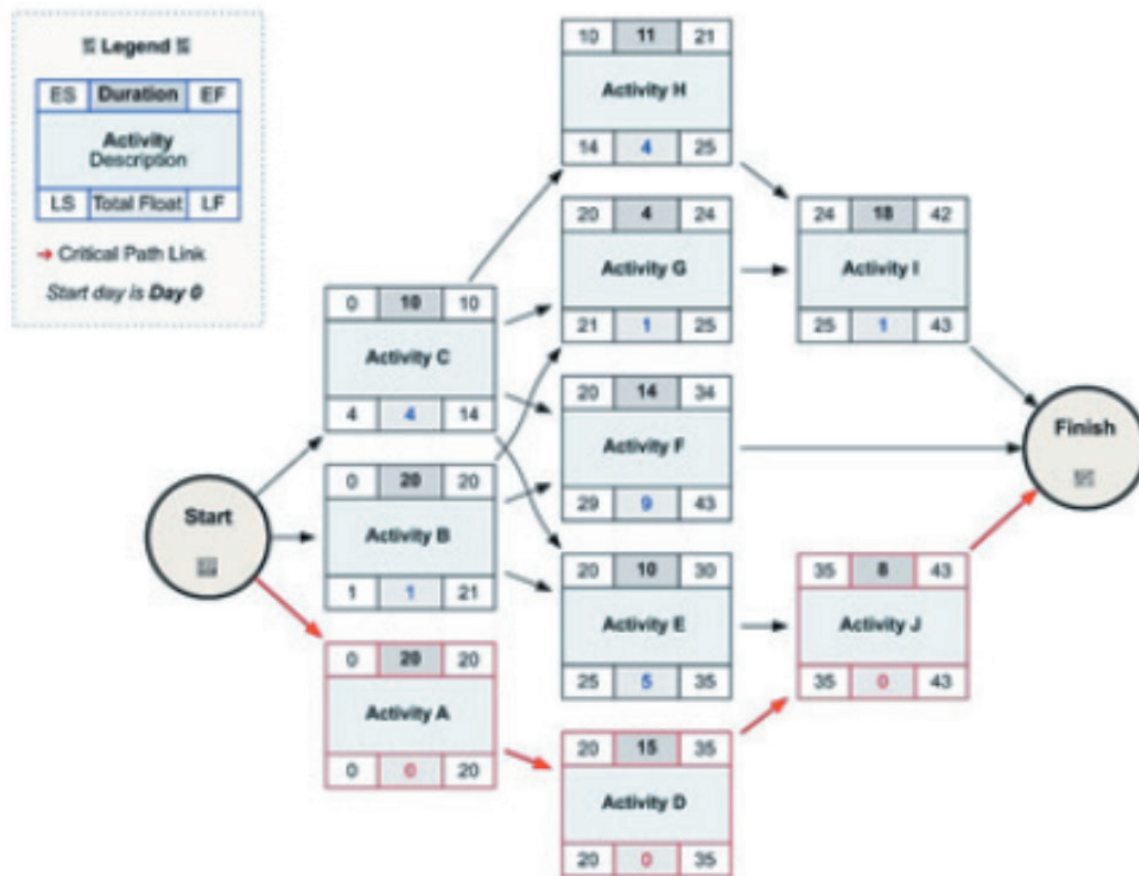


Figure 1. Activity-On-Arrow Diagram for selected project

Abbreviations: ES (Early Start), EF (Early Finish), LS (Late Start), LF (Late Finish)

The performance of both tools was evaluated based on the following criteria:

1. Accuracy of results – Comparison of simulated project durations (mean, variance, and probability distributions) between both tools.
2. Computational efficiency – Time taken to run simulations for each tool.
3. Usability – Assessment of user interface and ease of use, particularly focusing on web-based accessibility for the open-source tool.
4. Cost – A cost comparison between Oracle Crystal Ball and the PERT-based open-source tool, considering licensing fees and scalability.

4. RESULTS AND DISCUSSION

The histogram presented in Figure 2 provides the distribution of simulated project durations, while the S-curve represents the cumulative distribution function (CDF) obtained from this data. Above the chart, the legend is given to interpret bar heights and frequency of different critical paths. To further analyse the simulation results (e.g., in Excel), the tool enables the extraction of the raw generated data as CSV files for easy processing and creation of customized stacked histograms in Excel.

The following table provides the results on simulated durations, critical path and percentile probabilities for both software options. While Crystal Ball provided a slightly wider minimum and maximum range, both simulations resulted with similar average duration values. The critical path analysis resulted with similar results (more than 50%) for the main critical path (A>D>J), with slightly different frequency for the two alternative critical paths (B>G>I and C>H>I). Finally, both options provided practically identical results when it comes to percentile probabilities of project completion.

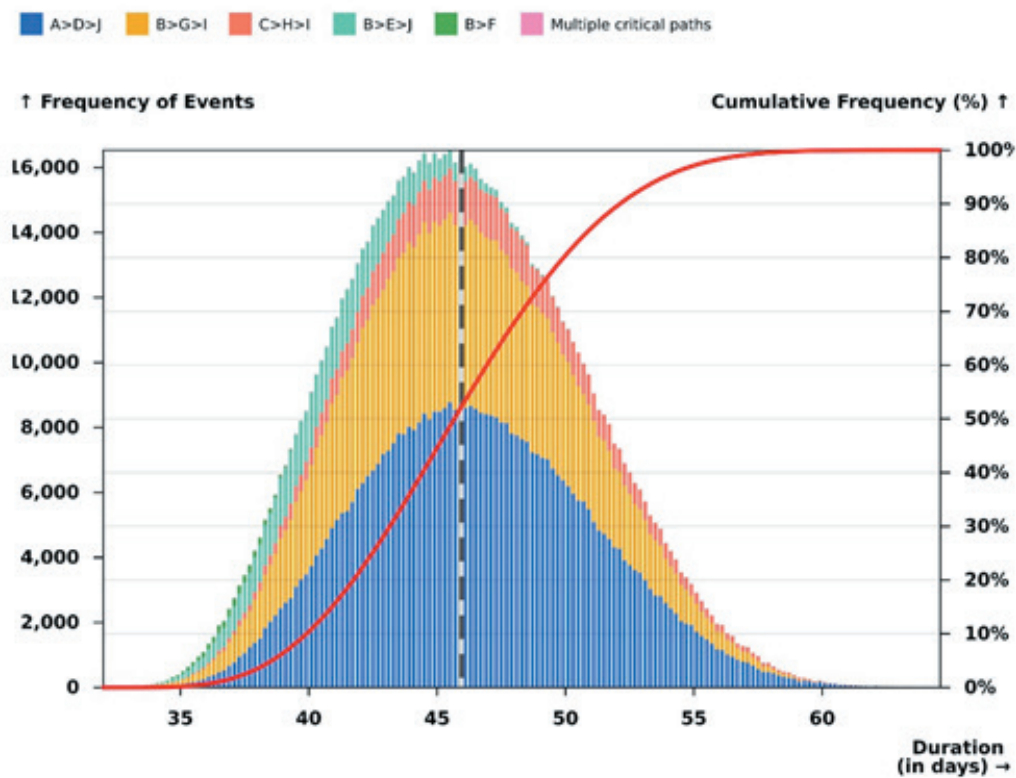


Figure 2. Open-source beta-PERT tool - Frequency chart for project completion time

Table 1. Summary statistics, critical path and percentile probabilities of project completion

	Summary	Oracle Crystal Ball	Open-Source PERT tool
Statistics	Minimum	31.79 days	33.04 days
	Maximum	64.69 days	60.52 days
	Average (μ)	45.95 days	45.89 days
	Standard deviation (σ)	4.60 days	5.17 days
	Variance (σ^2)	21.131 [days ²]	26.73 [days ²]
Critical Path	A>D>J	54.1%	50.7%
	B>G>I	23.6%	34.2%
	C>H>I	17.3%	8.3%
	B>E>J	4.8%	6.3%
	B>F	0.1%	0.4%
	Multiple critical paths	0.1%	0.1%
Percentile probability	50%	46 days	46 days
	60%	47 days	47 days
	70%	49 days	48 days
	80%	51 days	50 days
	90%	53 days	52 days
	100%	61 days	62 days



4.1. ACCURACY OF RESULTS

Both software tools produced highly accurate and comparable results in terms of the mean, variance, and probability distributions of project durations. The slight differences observed were negligible and within the expected tolerance, depending on random sampling. Before its release, the open-source tool was rigorously tested, including assessments with multiple probability distributions, to confirm its reliability. The normal and beta-PERT distributions are recommended, as they more accurately represent the natural variability in project timelines.

When using the normal distribution, activity durations might sometimes exceed the PERT bounds, which could lead to inconsistencies in the critical path list. This occurs because the random value generation method relies on a normal distribution calculated from the specified mean (μ) and standard deviation (σ), derived from PERT formulas. Although this method provides a close approximation of PERT distributions, it does not inherently respect the optimistic and pessimistic bounds defined by a PERT distribution.

4.2. COMPUTATIONAL EFFICIENCY

The computational time for both tools was found to be similar, with Oracle Crystal Ball taking slightly longer due to the overhead of running within the Microsoft Excel environment. The open-source tool, being web-based, was optimized for parallel processing and handled larger datasets efficiently. The simulation runs relatively quickly, but its speed depends on your computer's performance.

This web-based tool takes advantage of modern browser optimizations, offering notable benefits compared to Excel and Visual Basic for Applications (VBA). By utilizing JavaScript, it delivers faster computation, improved management of large datasets, and greater interactivity, resulting in a more efficient and adaptable solution. However, a user can still download the raw simulation data as a CSV file for further analysis in Excel or any other spreadsheet software.

4.3. USABILITY

The user interface of Oracle Crystal Ball, while robust, requires training in setting up a model and is less flexible when integrating with non-Excel environments. In contrast, the open-source tool's web-based interface is

intuitive and can be accessed from any device, requiring minimal setup. This flexibility makes the open-source tool an attractive option for users with limited technical expertise.

The open-source PERT tool has effectively simplified Monte Carlo schedule analysis by demonstrating how quickly a simulation can be executed in just a few seconds. Additionally, it automatically generates an activity-on-node diagram, further enhancing the efficiency and streamlining of the schedule analysis process (Figure 1).

The open-source nature enables further upgrades, customization and modifications for advanced users [16]. For those interested in exploring alternative distributions, the JavaScript file includes functions like $\text{normInv} = f(\text{mean}, \text{stdDev})$ for the Normal Distribution and $\text{betaPERT} = f(O, ML, P)$ for the beta-PERT Distribution. Additionally, users can create custom distribution functions tailored to their specific needs, all of which are called by the `runMonteCarloSimulation()` function [13].

4.4. COST

The most significant advantage of the open-source tool is its cost, which is zero. In contrast, Oracle Crystal Ball incurs significant licensing fees, which can become prohibitively expensive for smaller firms or individual users. Current prices for Oracle Crystal Ball vary between \$410 for university users, \$1,100 for the stand-alone installation on-premises, and 2,200 for the bundle installation with Crystal Ball Decision Optimizer.

5. CONCLUSION

This study demonstrates that the open-source Monte Carlo PERT-based simulation tool is a viable and effective alternative to traditional software like Oracle Crystal Ball. It provides similar accuracy, computational efficiency, and usability, with the added benefits of being cost-free and highly accessible. As project management increasingly shifts towards more cost-effective and flexible solutions, open-source Monte Carlo simulation tools hold great promise for the wider use of advanced risk analysis techniques in project scheduling.

Both tools produced highly accurate results, with negligible differences in mean, variance, and probability distributions. The open-source tool, which supports normal or beta-PERT distributions, underwent rigorous testing and effectively approximates PERT distributions, though the normal distribution may occasionally produce values outside PERT bounds.



In terms of computational efficiency, both tools performed similarly, but the open-source tool, being web-based and leveraging JavaScript, offered faster processing, better handling of large datasets, and enhanced interactivity compared to Oracle Crystal Ball, which experienced delays due to its Excel integration.

Usability and cost are key differentiators. Oracle Crystal Ball's interface, while robust, requires training and represents a less flexible option. In contrast, the open-source tool features an intuitive, web-based interface accessible from any device, making it user-friendly for those with limited technical expertise. Its open-source nature also allows for customization and upgrades, appealing to advanced users. Most notably, the open-source tool is free, which makes it a more accessible and attractive option, particularly for smaller firms or individual users.

REFERENCES

- [1] F. Kalateh and M. Kheiry, "A review of stochastic analysis of the seepage through earth dams with a focus on the application of Monte Carlo simulation", *Archives of Computational Methods in Engineering*, vol. 31, no. 1, pp. 47-72, 2024.
- [2] J. C. Chow, "Monte Carlo Simulations in Nanomedicine: Advancing Cancer Imaging and Therapy", *Nanomaterials*, vol. 15, no. 2, pp. 117, 2025.
- [3] P. S. Chen et al., "Using sample average approximation with Monte Carlo simulation to solve the stochastic outpatient appointment scheduling problem", *Communications in Statistics-Simulation and Computation*, pp. 1-18, 2025.
- [4] I. Al Hasan and A. Hammad, "A Systematic and Objective Framework for Evaluating Subcontractor Performance Using Monte Carlo Simulation Coupled with the Analytic Hierarchy Process and a Linear Additive Utility Model", *Buildings*, vol. 15, no. 3, pp. 390, 2025.
- [5] G. Avlijaš, "Examining the value of Monte Carlo simulation for project time management", *Management: Journal of Sustainable Business and Management Solutions in Emerging Economies*, vol. 24, issue 1, pp. 11-23, 2019.
- [6] Oracle. "Crystal Ball (simulation) (PDF)". Oracle.com. Accessed: March 1, 2025. [Online]. Available: <https://www.oracle.com/middleware/technologies/crystalball.html>.
- [7] C. Weishaar, "Predicting the Impact of Resource Delays on a Construction Project's Critical Path using Monte Carlo Simulation", MSc Thesis, University of Arkansas, Fayetteville, 2018.
- [8] G. Avlijaš, "Using Earned Value Management for More Sustainable Project Schedule Control", *Management: Journal of Sustainable Business and Management Solutions in Emerging Economies*, vol. 29, issue 2, pp. 1-12, 2024.
- [9] D. Salkeld, *Project Risk Analysis: Techniques for Forecasting Funding Requirements, Costs and Timescales*. CRC Press, 2016.
- [10] M. Vanhoucke, *Integrated Project Management Sourcebook: A Technical Guide to Project Scheduling, Risk and Control*. Springer, 2016.
- [11] R. Wanner, *Project risk management: the most important methods and tools for successful projects*. CreateSpace Independent Publishing Platform, 2013.
- [12] E.B. Malykh, "International Experience in the Risk Management of Investment Projects in the Field of Multi-Family Housing Using the Monte Carlo Method and the Oracle Crystal Ball Application for Microsoft Excel", *Economics and Management*, issue 11, pp. 92-97, 2018.
- [13] C. Yamahata. "PERT-Based Monte Carlo Schedule Analysis". Observablehq.com. Accessed: March 1, 2025. [Online]. Available: <https://observablehq.com/@christophe-yamahata>
- [14] W. East, *Critical Path Method (CPM) Tutor for Construction Planning and Scheduling*. New York: McGraw-Hill Education, 2015.
- [15] H. Kerzner, *Project Management: A Systems Approach to Planning, Scheduling, and Controlling*. Hoboken, NJ: Wiley, 2017.
- [16] C. Grisar and M. Meyer, "Use of Monte Carlo simulation: an empirical study of German, Austrian and Swiss controlling departments", *Journal of management control*, vol. 26, pp. 249-273, 2015.



MICROSOFT COPILOT AS A TRANSFORMATIVE TOOL IN BUSINESS: OPPORTUNITIES AND CHALLENGES

Milica Palić*,
[0009-0007-6810-0269]

Miloš Mravik,
[0000-0001-5442-3998]

Marko Šarac
[0000-0001-8241-2778]

Singidunum University,
Belgrade, Serbia

Abstract:

Artificial intelligence (AI) is progressing rapidly while impacting various industries, changing how businesses operate, and how employees perform tasks. Microsoft Copilot, an AI-powered assistant, is one of the most important developments, designed to assist employees with improving productivity by fostering collaboration, decision making and automation through the organization. This paper explores how Microsoft Copilot integrates and benefits enterprises, and shows challenges and implications for the employees and the organizations.

This study examines how Microsoft Copilot integrates into business workflows, especially in the Microsoft 365 applications, and it analyzes its impact on productivity and innovation. It also shows the key challenges, including resistance of the employees to adopting AI, security of data, and ethical considerations related to AI-assisted decision-making. Also, the paper analyses the changes in employment that require new skill adaptation as AI technology is broadly used in business settings.

This research assesses Microsoft Copilot and its impact on AI integration within Microsoft's ecosystem through a case study and empirical analysis. The findings can help companies in using AI while reducing risk and complexity. This paper also discusses the implications of AI in the future of work strategies, along with human capital deployment and business management.

Keywords:

Microsoft Copilot, Artificial Intelligence, Business Automation, Workforce Transformation, AI Adoption.

INTRODUCTION

Artificial intelligence (AI) is transforming industries' operations and how employees perform their tasks. Microsoft Copilot, as a productivity AI-augmented tool, is capable of automating business processes, forecasting effective decisions, and making daily tasks and work more efficient. This AI tool is integrated into Microsoft 365 business applications, and it uses generative AI to automate business processes.

In this paper, we explore the influence of Microsoft Copilot on modern enterprises, assessing its benefits, difficulties and overall impact on the workforce. However, the tool demonstrates significant efficiency improvements, it also comes with data security risks, requires employee adaptation, and raises ethical issues.

Correspondence:

Milica Palić

e-mail:

milica.palic.241@singimail.rs





Through the case study and empirical data, this paper will try to help in understanding how companies can strategically implement Microsoft Copilot while minimizing the potential risk. [1]

2. THE ROLE OF MICROSOFT COPILOT IN BUSINESS

Microsoft Copilot is designed for organizations to smoothly integrate with Microsoft 365 apps such as Word, Excel, Outlook, and Teams. Employees can use it for automating repetitive tasks, collaborating with their teams and creating content faster. According to industry reports, 77% of users said that they didn't want to stop using Microsoft Copilot after trying it. [2]

2.1. KEY BENEFITS OF MICROSOFT COPILOT

Productivity and efficiency in Copilot can help in reducing the time spent on manual tasks, so employees can focus on higher-value work. Research shows that 64% of employees struggle with energy and time, while 70% delegate their tasks to AI. [3] [4] Copilot solves these challenges by:

- Real-time recommendations for data analysis and reporting.
- Automating workflows in PowerPoint and Excel. Generating content in Word, Teams, and Outlook.

Industry-specific and fostering creativity in Copilot, in addition to automating repetitive tasks, can also foster creativity, help with content creation, and assist with building strategies. As an industry-oriented it can help with:

- Finance – Help with reporting and data validation.
- Sales – Help with lead prioritization.
- Customer Service – Help with monitoring of issues and faster answers.
- HR – Simplify paperwork and assist with hiring.

2.2. KEY CHALLENGES OF AI IN BUSINESS

Microsoft Copilot and other AI tools have a lot of benefits and opportunities in the business world, but there are also a lot of challenges for their adaptation. In the real world, to benefit from AI, companies need to overcome key challenges such as data security concerns, employee resistance, complexity, and integration difficulties. [5]

1. Employee resistance and lack of skills

One of the biggest concerns for employees is whether AI will drastically change or replace their jobs in the future, and that is the first obstacle to adaptation. According to studies, 82% of company leaders believe that employees would need new skills for AI adaptation. [6] [7]

To have a smoother transition, companies should:

- Educate employees that AI is meant to supplement human work, not to replace it.
- Invest in education and training initiatives to gain new skills required for AI adaptation in daily work.
- Demonstrate to employees on real initiatives and show them how Copilot can increase productivity and help with creativity and innovation.

2. Data security and privacy

Microsoft Copilot, like other AI tools, is processing a lot of business data daily. Besides that, Microsoft Copilot runs under the Microsoft 365 security framework, companies have a lot of concerns about data security and privacy, and they need to implement additional security measures to prevent:

- Access to data by unauthorized employees.
- Potential errors and bias in AI-generated content.
- Non-compliance with data protection standards such as GDPR.

Companies should implement and establish AI governance to reduce these potential risks:

- Controlled access and the ability to use AI insights.
- Strict and clear rules for using AI with sensitive personal data.
- Clear procedures for managing security threats and mitigating risks.

3. Potential compliance and ethical issues in business

The popularity of AI technology in business also increases ethical concerns for all companies. Data confidentiality, accountability, and transparency stand out as the biggest problems. Artificial intelligence models, especially those used in Microsoft Copilot, can inadvertently create problems with the accuracy of the data they deliver to the end user. One very important segment is that organizations need to properly manage risks, as well as consider local and global market regulations.



To improve business and avoid potential concerns with the use of AI technologies, companies should:

- Introduce modern mechanisms whose role is to verify the reliability and accuracy of the data obtained by artificial intelligence.
- Create ethical guidelines internally related to artificial intelligence. This way, companies improve the quality of decisions made.
- Harmonize artificial intelligence systems with regulatory authorities in the country and around the world. By harmonizing systems, companies reduce legal risks that can potentially cause very significant damage.

4. Technical challenges of company systems

A major challenge for all companies, regardless of whether they are small or large, is organizing and managing the integration of artificial intelligence into current operations. By analyzing companies that have begun the process of introducing AI technologies into their systems, we conclude that 75.2% of companies are struggling with the integration of artificial intelligence. Some of the problems are:

- The existing infrastructure is outdated and not ready or compatible with the process automation offered by artificial intelligence.
- Staffing problem, lack of education of engineers.
- The problem of company management is not trendy and does not follow the world of artificial intelligence.

For companies to overcome these challenges, it is necessary to conduct a detailed analysis of the readiness of the entire infrastructure. After completing the analysis, it is necessary to establish plans and phases for incorporating artificial intelligence into current operations.

5. Cost and ROI considerations

Although artificial intelligence is expected to reduce costs and increase system efficiency, companies must carefully create plans for financing such projects.

Based on information obtained from many companies that have introduced artificial intelligence into their systems, it is estimated that the expected return on investment is \$3.5 for every dollar invested. The following items also affect ROI:

- Quality of the strategy plan.
- Detailed education of employees who will use AI tools daily.
- Continuous improvement of AI systems to improve the current system.

One way to measure success with the introduction of artificial intelligence is to establish KPIs. Their role is to monitor current productivity, efficiency, and cost reduction.

6. Change management and user adoption

The introduction of artificial intelligence is changing company culture. Many organizations make mistakes by making incorrect assessments about the successful integration of AI systems within their business. Key obstacles include the following:

- Lack of trust in artificial intelligence. Employees spend too much time relying on artificial intelligence systems.
- 56% of employees report having problems adopting new technologies. The problems cited include finding the time and energy to learn.
- 64% of employees use artificial intelligence tools on their accounts instead of their business accounts. This compromises data security.

For organizations to successfully adopt artificial intelligence systems, they need to:

- Provide quality and professional training to all employees.
- Find employees who are best at using these tools and appoint them as team leaders whose role will be to educate other employees.
- Improve employee culture regarding artificial intelligence systems.

3. THE FUTURE OF AI IN BUSINESS

In parallel with the development of artificial intelligence and the introduction of Microsoft Copilot, companies must create strategic approaches to ensure the highest quality integration. The following key steps stand out:

- Understanding the current system. Companies need to assess their current business and, based on the results obtained, find areas where AI can provide the most value.
- Establishing proper policies for the ethical use of artificial intelligence. It is necessary to ensure data protection, which is one of the most important items for maintaining trust and compliance in companies.



4. CASE STUDY – USE OF MICROSOFT COPILOT FOR 4 MONTHS

To write this scientific paper, a case study was conducted on an employee working in the IT industry. The case study presented the employee's daily activities. The results obtained were very good and confirmed that the use of artificial intelligence in the right way can greatly facilitate various daily activities.

The key factors used to write this scientific paper are:

- The Microsoft Copilot tool was used. The testing lasted 4 months. The data included are those collected during working weeks (Monday-Friday).
- Microsoft Copilot was used only for basic daily tasks. Information or documentation was searched through Copilot chat. A segment of composing email or meeting summaries, making various corrections was also included. The list of software used is Microsoft Excel, Microsoft Outlook, Microsoft PowerPoint, Microsoft Teams, and Microsoft Word.
- Three metrics are tracked:
 - Time Saved - Obtaining an estimate of how long a task would take to solve without the use of artificial intelligence.
 - Time Spent - Obtaining an estimate of how long it took to solve a task using artificial intelligence.

- Time Gained - The difference between time saved and time spent (Time Spent - Time Saved).

To achieve the best possible results, negative values have also been introduced. In the case where Microsoft Copilot was unable to help, a negative value is recorded. In this case, the result obtained gives us information that more time was spent than originally estimated. To correct grammatical errors and email structure, the maximum time is 1 minute. In this case, the tasks improved the result because Microsoft Copilot improved the overall quality of the text. The tested candidate was not a native English speaker. While using Microsoft Copilot, the candidate improved the clarity of his writing all the time.

4.1. GRAPHICAL DISPLAY OF THE RESULTS OBTAINED

Table 1 shows information about Copilot Tasks.

Figure 1 shows information about the Number of Prompts per Day.

Figure 2 shows information about the Sum of gained time in minutes.

Table 1. Copilot Tasks

Application	Task	Sum of Time (minutes)
Copilot Chat	Searching for information	60
	Composing a Summary of bullet points	44
	Correcting Text	25
	Composing the Agenda for the events	15
	Correcting Emails	12
	Searching Inbox	9
	Searching for people in the org.	8
	Chatting with Copilot	0
	Searching for a PPT File	-4
	Searching for documents	-5
Copilot Chat Total		164
Excel	Creating visuals out of the Excel Table	20
Excel Total		20



Application	Task	Sum of Time (minutes)
Outlook		62
	Outlook Total	62
PowerPoint	Creating Slides from the Word document	50
	Creating Footnotes from a PPT presentation	35
	Arranging the Presentation	30
	PowerPoint Total	115
Teams Meetings	Meeting Recap	427
	Meeting Catch-up	25
	Creating notes from the meeting	0
	Teams Meetings Total	452
Word	Composing a Word document out of a website with comments	55
	Word Total	55
	Grand Total	868

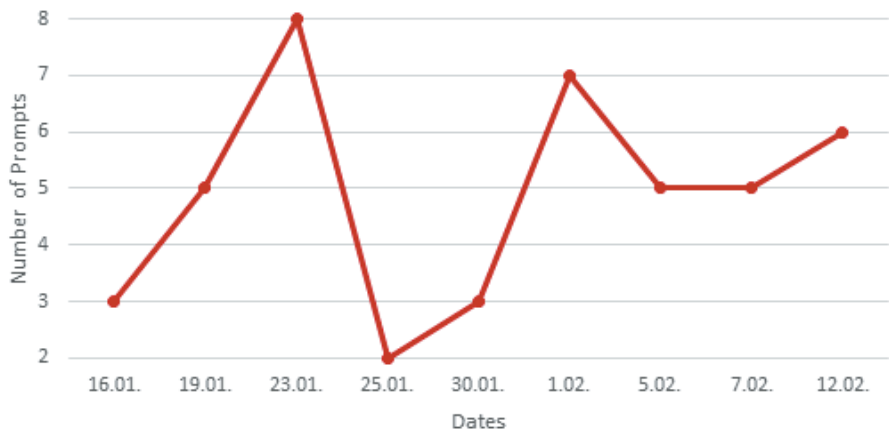


Figure 1. Number of Prompts per Day

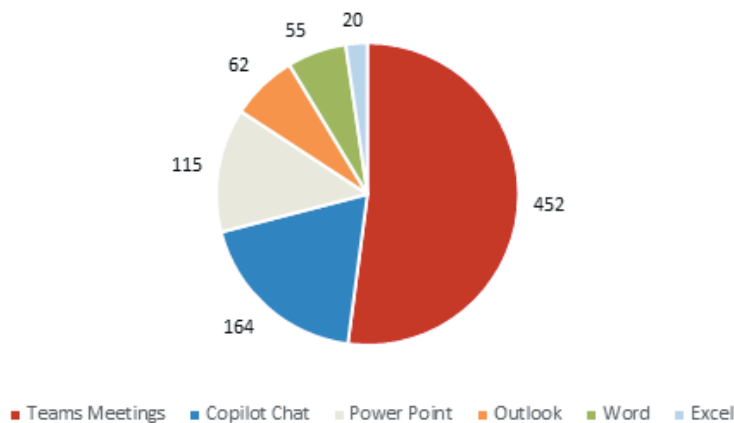


Figure 2. Sum of gained time (minutes)



4.2. KEY FINDINGS AFTER THE CASE STUDY

- Microsoft Copilot gave excellent results in most cases. Based on the results, Copilot was successful in 92.54% of the tasks received. In 41.26% of cases, it gave better results than the user himself. In this case, not only is time saved, but the quality of work is also improved.
- Microsoft Copilot performed very well in advanced document searches.
- The biggest time savings were shown within the meeting recap. The user did not need to watch the entire video. In this case, a comparison was made with watching the video at 2x speed. Examples: if the meeting lasted 90 minutes, it took 15 minutes to view the summary. In this case, 30 minutes were saved. In case of overlapping meetings, this option is very useful.
- During the testing period, the candidate did not create many Word and PowerPoint documents. The impact of Copilot on these applications is not fully reflected.
- The candidate who was tested is very proficient in Microsoft Excel software and the results showed that within that software, Microsoft Copilot was not significantly useful.

The following limitations existed during testing:

- At the beginning of the testing, the candidate worked only in Microsoft Excel Online.
- The candidate worked only within an Excel spreadsheet.
- The candidate could not process large amounts of data (20,000+ rows).

During the testing period, Microsoft Copilot saved more than 14 hours of active work (almost two full working days). As already noted, the meeting recap option gave the best results. The results obtained highlight the potential of artificial intelligence and Microsoft Copilot, which directly affects quality improvement, process simplification, and work quality improvement. [8]

5. CONCLUSION

To 'survive' in today's advanced world, and progress together with technology, we must accept that AI is no longer an option, but something that will follow us everywhere. Microsoft Copilot is an AI business tool that provides an excellent example of how AI can transform and facilitate decision-making, productivity and creativity as well as teamwork and collaboration. However, as we described, the adoption of such tools is one of the most difficult barriers that companies need to overcome. Data security concerns, as well as employee resistance and their fear that AI tools may replace their positions in the future, are among the main barriers to adoption. [9]

Key findings of this study show what the advantages and opportunities are, as well as the obstacles that companies face when implementing solutions based on artificial intelligence. The most important steps for the successful integration of AI tools are, first the assessment of the company, based on which a well-planned implementation approach will be made that includes training of personnel as well as clear policies of artificial intelligence. The study clearly shows us through the results that AI-based tools can significantly save time and improve efficiency, which primarily depends on how well it is aligned with user needs and existing business flows. [10]

What is clear to all of us is that companies that strategically embrace and follow the advancement of AI technology have a competitive advantage over companies that avoid it.

Companies must proactively adapt to the advancement of AI-based tools to be better equipped to develop and navigate in a changing business environment and not fall behind as the business evolves.

6. ACKNOWLEDGEMENTS

We would like to acknowledge the company Crayon for their support in providing us with resources and licenses so we can continue our development and research in this field.



REFERENCES

- [1] F. O. Olatoye, K. F. Awonuga, N. Z. Mhlongo, C. V. Ibeh, O. A. Elufioye and N. L. Ndubuisi, "AI and ethics in business: A comprehensive review of responsible AI practices and corporate responsibility," *International Journal of Science and Research Archive*, vol. 11, no. 1, pp. 1433-1443, 2024. doi: 10.30574/ijrsra.2024.11.1.0235.
- [2] B. Wiseman, "What Can Copilot's Earliest Users Teach Us About Generative AI at Work?," 2023. [Online]. Available: <https://www.microsoft.com/en-us/worklab/work-trend-index/copilots-earliest-users-teach-us-about-generative-ai-at-work>. [Accessed 31 03 2025].
- [3] P. Hemmer, M. Westphal, M. Schemmer, S. Vetter, M. Vössing and G. Satzger, "Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction," in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, Sydney, 2023. doi: 10.48550/arXiv.2303.09224.
- [4] C. Vasilescu and M. Gheorghe, "Improving the Performance of Corporate Employees through the Use of Artificial Intelligence: The Case of Copilot Application," in *Proceedings of the International Conference on Business Excellence*, Warsaw, 2024. doi: 10.2478/picbe-2024-0153.
- [5] G. Ismail, R. Paavo, A. Ahmad, M. Brad and A. Imran, "Confronting and alleviating AI resistance in the workplace: An integrative review and a process framework," *Human Resource Management Review*, vol. 35, no. 2, p. 101075, 2025. doi: 10.1016/j.hrmr.2024.101075.
- [6] L. Zuowei, "AI Ethics and Transparency in Operations Management: How Governance Mechanisms Can Reduce Data Bias and Privacy Risks," *Journal of Applied Economics and Policy Studies*, vol. 13, no. 1, pp. 89-93, 2024. doi: 10.54254/2977-5701/13/2024130.
- [7] C. Ikegwu, N. S. Uzougbo and A. O. Adewusi, "Legal accountability and ethical considerations of AI in financial services," *GSC Advanced Research and Reviews*, vol. 19, no. 2, pp. 130-142, 2024. doi: 10.30574/gscarr.2024.19.2.0171.
- [8] L. Perne, "How I saved more than 14 hours in a month by using Copilot for M365!," 2024. [Online]. Available: <https://www.linkedin.com/pulse/how-i-saved-more-than-14-hours-month-using-copilot-m365-luka-perne-yfwgf/>. [Accessed 31 03 2025].
- [9] M. Mravik, T. Vetrisevi, K. Venkatachalam, M. Sarac and N. Bacanin, "Diabetes Prediction Algorithm Using Recursive Ridge Regression L2," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 457-471, 2022. doi: 10.32604/cmc.2022.020687.
- [10] M. Šarac, M. Mravik, J. Dijana, Š. Ivana, Ž. Miodrag and B. D. Nebojša, "Intelligent diagnosis of coronavirus with computed tomography images using a deep learning model," *Journal of Electronic Imaging*, vol. 32, no. 2, pp. 1-10, 2022. doi: 10.1117/1.JEI.32.2.021406.



IDENTIFICATION OF SAFE HELICOPTER LANDING ZONES AND AN OVERVIEW OF HELIPORTS IN SERBIA

Zoran Kričković^{1*},
[0009-0006-1509-5008]

Emina Kričković²
[0000-0002-0737-5636]

¹Military Geographical Institute –
“General Stevan Bošković”,
Belgrade, Serbia

²Faculty of Geography,
University of Belgrade,
Belgrade, Serbia

Abstract:

Helicopters are most commonly used in inaccessible areas during various operations, whether civilian or military. This paper aims to provide an overview of existing research on the automatic generation of safe helicopter landing zones and to define the necessary parameters for their application in the territory of the Republic of Serbia. The paper also presents the present-day state of designated helicopter landing areas in Serbia. Notably, with 12 heliports, Serbia is among the leading countries in the region for heliport infrastructure. Additionally, the paper outlines the legal frameworks governing this subject in Serbia, the European Union, and the United States. Finally, the paper proposes parameters that define safe helicopter landing zones. Research of this kind could aid decision-making processes during emergencies in the Republic of Serbia and other countries.

Keywords:

Helicopter, Safe Landing Zone, Heliport, Terrain Slope.

INTRODUCTION

Helicopters are most commonly used in inaccessible areas during various operations, whether civilian or military. With the rapid development of unmanned aerial vehicle (UAV) technology for the same purposes, there is an increasing need for geographic data on these inaccessible regions.

Over the past two decades, the demand for real-time digital surface model generation has grown, particularly for aircraft landings under low visibility conditions [1]. In recent years, with advancements in UAVs and artificial intelligence [2], the need for suitable data for identifying surfaces suitable for vertical take-off and landing (VTOL) aircraft has also increased.

One of the key topics for the armed forces is the assertion by [3] that the ability to rapidly conduct precise terrain assessments and remotely detect associated threats represents a key priority for the United States

Correspondence:

Zoran Kričković

e-mail:

zoran.krickovic@vs.rs





Armed Forces. In [4] stated that UAVs were initially used primarily for military purposes, rescue operations, and disaster response, but have since been widely adopted for civilian and commercial applications.

This paper aims to provide an overview of existing research on the automatic generation of safe helicopter landing zones, define the necessary parameters for their application in the Republic of Serbia, and present the regulatory framework for heliports. Additionally, the paper examines the present-day state of designated helicopter landing areas in Serbia.

2. LITERARY REVIEW

In [5] state that for helicopter landings in inaccessible areas, such as Yosemite National Park in the United States, a flat terrain free of tree canopies and other hazards is required. Landing a helicopter on the surface and transporting the injured provides an optimal procedure, whereas other techniques, such as hovering to extract an injured person, increase the risk during rescue operations [5].

Currently, for identifying surfaces suitable for vertical take-off and landing (VTOL) aircraft, the most popular method is the processing of 3D LiDAR (Light Detection and Ranging) point clouds collected by LiDAR technology mounted on an aeroplane or helicopter [1]. This claim is supported by numerous articles on the subject. Despite this advanced technology, maps are still necessary for additional analysis. Authors in [1] emphasized the importance of comparing LiDAR collected data with map data to enable pilots to identify the safest landing zones for helicopters [1]. This topic is not only a priority for military applications but also for civilian use, as demonstrated by [6], who similarly used topographic maps to assess obstacles and identify the most suitable helicopter landing zones near high-risk forested areas due to wildfires in Golestan National Park, north-eastern Iran.

According to [6], state that firefighting using helicopters is a common global practice. They further note that helicopters are primarily used in rural areas with underdeveloped road networks. They identified a helicopter landing zone as a flat area devoid of trees, shrubs, logs, and large rocks. In wildfire suppression over unknown forested terrain, it is crucial to assess the safety and efficiency of landing sites [6]. A landing site should be free of obstacles and vegetation within a 50–75 m diameter. Additionally, the terrain slope at the landing site should be less than 5%. In this study, the authors

first created a wildfire risk map by collecting data from various sources, including vegetation cover type, physical geographical features, climate, and human factors [6]. To identify safe helicopter landing zones, they defined a minimum area of 50 meters in diameter (greater than 2,000 square meters) alongside other conditions critical for proximity to firefighting areas [6]. The slope analysis was conducted using a Digital Elevation Model (DEM) and a Triangulated Irregular Network (TIN). To identify clear spaces, they used high-resolution satellite images from Google Earth [6].

Authors in [5] compared two GIS-based scientific methods for identifying safe helicopter landing zones using available GIS data in Yosemite National Park, USA. The first method relied on expert judgment, while the second employed machine learning. They concluded that both methods yielded similar results. In both approaches, safe landing zones were identified as areas with a slope of less than 5% and a 90-meter radius free of natural or human-made obstacles [5].

Similarly, [1] stated that real-time identification of safe helicopter landing zones has a significant impact on saving lives, reducing risk, and improving operational efficiency, particularly in search and rescue or medical transport missions, where immediate and accurate determination of a safe landing zone can mean the difference between life and death. In the same study, the authors argued that real-time identification of safe helicopter landing zones in urban environments is highly valuable due to the challenges posed by various obstacles, such as buildings and other structures that frequently change [1]. Additionally, they emphasized that situational awareness and spatial perception are crucial during landings under adverse meteorological conditions, such as fog, smoke, darkness, or sandstorms, which impair helicopter manoeuvrability [1]. A major limitation in real-time safe zone identification is the complexity of computational algorithms, which require extensive processing power and must be significantly simplified for practical real-time use [1]. The approach employed in this study involved first processing LiDAR-generated point clouds, and then creating a slope and roughness map based on the scanned terrain. The pre-defined slope and roughness values were applied to the resulting map to identify safe landing zones. The threshold values in this study, as defined by the authors, were 24×24 meters for the landing area and a terrain slope of 4%, which aligns with established helicopter landing standards [1].



The application of artificial intelligence (AI) in UAV image recognition for rescue missions has been investigated by [2]. AI was utilized for deep learning in target recognition algorithms. The target depends on the UAV's specific purpose. Unlike traditional target recognition algorithms, deep learning-based algorithms can autonomously learn feature representations and detection models necessary for identifying targets from large volumes of raw data, eliminating the need for manually designed features [2]. According to [2], as early as 1997, researchers at Cornell University proposed an automatic target recognition model for UAVs.

Today, UAV usage in military conflicts has become widespread, as evidenced by footage shared on social media platforms from conflicts in Afghanistan, Armenia, Russia, and Ukraine. For navigation, [2] note that UAVs utilize GPS, LiDAR, inertial navigation systems, and visual navigation systems. They highlight both the advantages and disadvantages of GPS, such as the increased risk of signal loss and potential electromagnetic interference in mid-latitude regions with weak signals [2]. The visual navigation system primarily involves measuring various navigation parameters through images captured from the ground using onboard imaging equipment (e.g., lighting systems, infrared imagery, Synthetic Aperture Radar (SAR), etc.) [2]. In [2] state that visual sensors not only facilitate efficient localization and mapping but also detect targets and extract effective spatial semantic information, making them particularly useful for guiding electric UAVs in search and rescue operations.

Geospatial software designed for identifying military helicopter landing zones in extremely challenging environments with minimal terrain preparation is presented in [3]. Their research demonstrated that a DEM with a 30-meter resolution identified numerous favourable landing zones, whereas a 1-meter resolution DEM revealed poor landing areas that the 30-meter resolution DEM failed to detect [3]. Furthermore, the study found that a 1-meter resolution DEM was overly precise, leading to the conclusion that a 5-meter resolution DEM is optimal. This research highlights the efficiency and effectiveness of using DEMs to rapidly identify safe helicopter landing zones [3].

3. HELICOPTER LANDING ZONES – HELIPORTS IN SERBIA

Areas designated and intended for helicopter landing are referred to as heliports. In Serbia, they are defined by the Regulation on Conditions and Procedures for Issuing Permits for the Use of Heliports. This subject falls under the jurisdiction of civil aviation, which in Serbia is managed by the Civil Aviation Directorate on behalf of the Government of the Republic of Serbia, operating as a public agency (hereinafter: the Directorate).

According to the Aviation Terminology Glossary of the Directorate, an airport is defined as an area (including all facilities, installations, and equipment) on land, water, or a fixed, coastal, or floating structure, intended wholly or partially for the landing, take-off, and movement of aircraft [7]. The same glossary defines a public-use airport/heliport as an airport or heliport accessible to all users under the same conditions, with its operational hours published in the Integrated Aeronautical Information Package [7].

For comparison, civil aviation in the United States is regulated by the Federal Aviation Administration (FAA), operating under the authority of the U.S. Department of Transportation. In the European Union, civil aviation falls under the jurisdiction of the European Union Aviation Safety Agency (EASA), established in 2002 as an independent and neutral body of the European Union [8].

According to the records of the Directorate [9], there are currently 12 heliports in Serbia, with their locations presented in Figure 1. As clearly depicted in the figure, the distribution of heliports across Serbia is highly uneven.

The southwestern region of Serbia, particularly Zlatibor County, has the highest coverage, with four heliports. In contrast, the entire eastern and north-eastern regions of the country lack any heliport infrastructure. Notably, the Stara Planina mountain region, a well-known tourist destination, is entirely without heliports. This absence of heliport facilities could limit the tourism potential of the area by restricting emergency medical access and reducing the overall accessibility for visitors.

Table 1 presents a comprehensive list of all heliports in Serbia. The data indicate that only seven administrative counties, out of 30 in the Republic of Serbia, are equipped with heliports. Furthermore, analysing the distribution at the municipal level, it is evident that only 12 out of 197 municipalities in Serbia have at least one heliport. This limited coverage highlights significant regional disparities in heliport accessibility, which could implicate emergency response, transportation, and regional development.

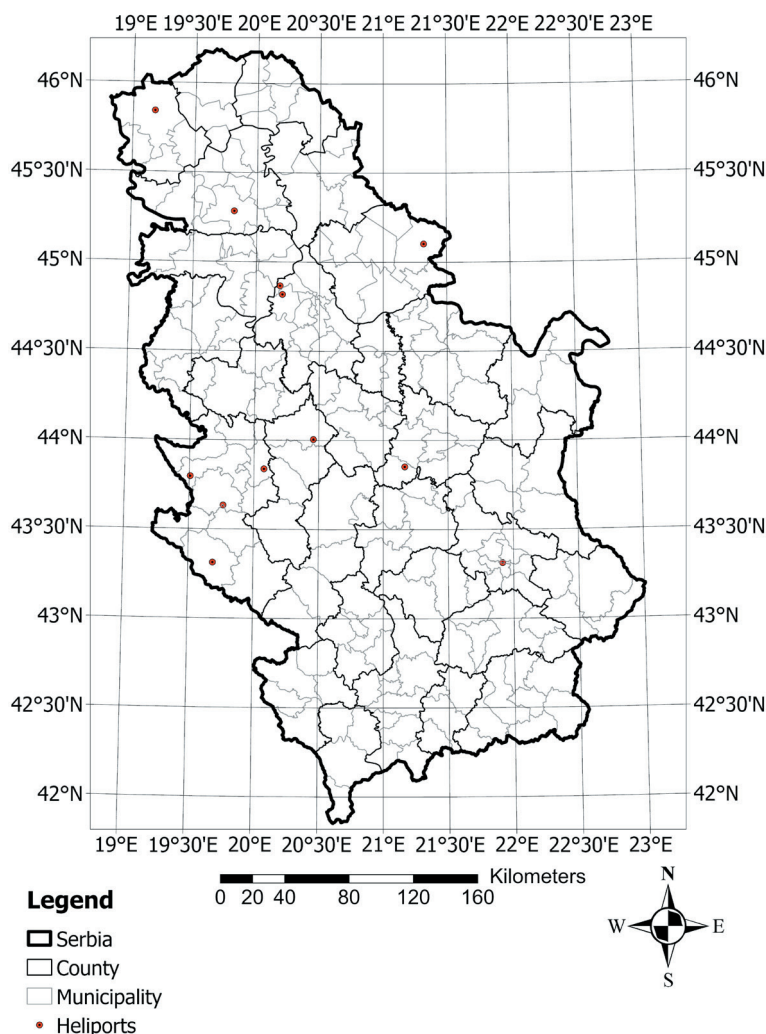


Figure 1. Heliports location in the Republic of Serbia. Source: authors, [9]

Table 1. List of Heliports in Serbia According to the Records of the Directorate of Civil Aviation of the Republic of Serbia [9]

Heliport name	Coordinates (WGS84)		Municipality	County	Permit date
	N	E			
Mokra Gora	434742.92	0193027.72	Užice	Zlatiborski	20.01.2016.
Novi Sad Ciklonizacija	451703.16	0194909.40	Novi Sad	Južnobački	24.02.2016.
BD Agro	444907.08	0201224.28	Surčin	Beogradski	21.07.2017.
Gornji Milanovac Takovo	440021.59	0202728.49	Gornji Milanovac	Moravički	20.11.2018.
Ljubiš Zlatibor	433802.00	0194552.00	Nova Varoš	Zlatiborski	07.06.2019.
Niš Klinički centar	431848.04	0215457.75	Medijana/Niš	Nišavski	08.11.2019.
Elgra Vižn	445201.65	0201115.18	Zemun	Beogradski	26.02.2021.
Ločika	435112.70	0210959.45	Rekovac	Pomoravski	19.08.2021.
Kalos	455029.68	0191031.64	Sombor	Zapadnobački	02.06.2022.
Požega	435021.28	0200437.75	Požega	Zlatiborski	31.10.2022.
Swisslion – Takovo Vršac	450613.13	0211906.74	Vršac	Južnobanatski	21.12.2022.
Lučice	431845.50	0194122.45	Prijepolje	Zlatiborski	19.10.2023.



Figure 2a presents an example of a heliport in Serbia, located in Mokra Gora, a popular tourist destination in western Serbia, within the Zlatibor County, Užice Municipality. The satellite image showing its exact location is shown in Figure 2b.

According to [11], Serbia had eight heliports in 2024. However, data from the Civil Aviation Directorate of Serbia indicate a total of 12 heliports. Also, according to the Plan for Organizing the Search and Rescue System for Aircraft and Persons in Civil Aviation of the Republic of Serbia [12], the primary participants in search and rescue operations are the Ministry of Internal Affairs, the Ministry of Defence, the Ministry of Health (specifically the Clinical Centres of Serbia and Niš), the Mountain Rescue Service Association, and the Red Cross of Serbia. This ensures that the capacities of all these ministries and organizations can be utilized during such operations, as the Directorate has signed cooperation agreements with them.

Secondary participants in search and rescue operations include all entities listed in the Regulation on the Engagement of Participants in Aircraft Search and Rescue in Civil Aviation [13], with whom the Directorate has not signed cooperation agreements and whose engagement is not specifically defined by law. Secondary participants may include authorities of autonomous provinces and local self-government units, public enterprises, and all-natural and legal persons capable of assisting, as well as aero clubs, aviation organizations, aviation associations, and specialized rescue units [12].

4. TERRAIN CONDITIONS DEFINING A FAVOURABLE SURFACE FOR HELICOPTER LANDING

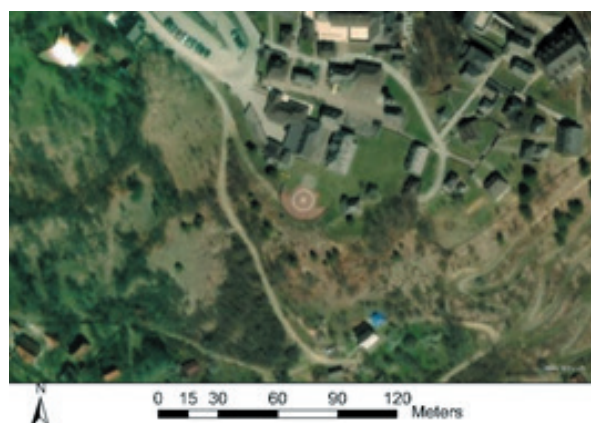
As previously mentioned, in addition to modern technologies used to determine safe landing sites, the most commonly utilized resource is the DEM, which is the basis for calculating terrain slope. Terrain slope is one of the most critical parameters for identifying safe helicopter landing sites. According to the Rule Book on the Conditions and Procedure for Issuing Permits for the Use of Heliports (hereafter referred to as the "Rule Book"), the terrain slope should not exceed 3% to 7%, depending on the helicopter classification [14].

According to Kumar, all single-engine helicopters belong to Class III, while Classes I, II, and III indicate a helicopter's capability to land safely in the event of engine failure [15]. In several cited studies, authors have defined the minimum required terrain slope for a safe helicopter landing zone as ranging from 4% to 5%.

The Rule Book also specifies that the minimum dimensions for helipads must comply with the helicopter manufacturer's manual [3]. Authors in [3] tested their software using the shortest (340 m) and narrowest (45 m) dimensions within their designated test area, covering approximately 15,300 square meters [3]. In [6] defined a safe landing zone as a minimum of 2,000 square meters, with a minimum radius of 50 meters, while [1] specified dimensions of 24×24 meters, totalling 576 square meters. The largest helicopter in the world measures 40×32 meters, and when compared to the previously mentioned parameters, the required area for a safe landing zone can range from 576 to 15,300 square meters.



a)



b)

Figure 2. View of the Mokra Gora Heliport on: a) an Aerial Photograph, b) a Satellite Image

Source: a) [10], b) authors.



Beyond these two parameters, soil type is also a crucial factor [3]. In [3] incorporated the National Land Cover Database (NLCD) from the U.S. Geological Survey (USGS), which has a 30-meter resolution, to assess terrain suitability. One of the main limitations in identifying safe landing zones is the availability of digital data required for analysis.

5. CONCLUSION

As presented in this study, LiDAR technology is predominantly utilized for assessing safe landing zones for helicopters or UAVs. Through terrain scanning and point cloud generation, LiDAR enables the creation of a digital surface model (DSM), which is subsequently used to produce slope maps. By correlating these maps with predefined parameters for safe landing zones, based on the size of the aircraft it is possible to automatically and proactively determine the most suitable landing areas in emergencies. The only distinction lies in whether this process is conducted in advance by experts or if AI and machine learning algorithms are applied in real time to identify optimal landing locations based on predefined criteria.

Helicopter or UAV landings have proven to be the safest method for rescuing injured individuals during emergencies. Therefore, the identification and definition of safe landing zones are necessary procedures in crisis management. Serbia is not lagging in legal regulations governing the use of helicopters and UAVs in such scenarios. However, it still falls behind more advanced nations in terms of the operational use of helicopters and UAVs for search and rescue and medical evacuation operations.

This research has demonstrated that DEM and slope maps are indispensable tools for defining safe landing zones for helicopters and UAVs. At the same time, large-scale topographic maps remain crucial for such analyses. Future research should explore recognized methodologies for generating slope maps from DEMs and investigate ways to integrate these datasets with digital topographic maps for enhanced accuracy.

Beyond spatial and cartographic data, future studies should also incorporate soil type classification to assess potential hazardous landing conditions despite favourable terrain parameters. This highlights the need for a multidisciplinary approach, requiring collaboration among experts from various fields to ensure comprehensive and reliable results.

6. ACKNOWLEDGEMENTS

This study was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grant No 451-03-136/2025-03/200091).

REFERENCES

- [1] A. Massoud, A. Fahmy, U. Iqbal, S. Givigi and A. Noureldin, "Real-Time Safe Landing Zone Identification Based on Airborne LiDAR," *Sensors*, vol. 23, no. 3491, pp. 1-26, 2023.
- [2] P. Tang, J. Li and H. Sun, "A Review of Electric UAV Visual Detection and Navigation Technologies for Emergency Rescue Missions," *Sustainability*, vol. 16, no. 2105, pp. 1-22, 2024.
- [3] S. N. Sinclair and S. A. Shoop, "Automated detection of austere entry landing zones: A "GRAIL Tools" validation assessment," *Transactions in GIS*, vol. 23, no. 6, p. 1317-1331, 2019.
- [4] J. A. Subramanian, V. S. Asirvadam, S. A. B. M. Zulkifli, N. S. S. Singh, N. Shanthi and R. K. Lagisetty, "Target Localization for Autonomous Landing Site Detection: A Review and Preliminary Result with Static Image Photogrammetry," *Drones*, vol. 7, no. 8, pp. 1-23, 2023.
- [5] P. Doherty, Q. Guo, and O. Alvarez, "Expert versus Machine: A Comparison of Two Suitability Models for Emergency Helicopter Landing Areas in Yosemite National Park," *The Professional Geographer*, vol. 65, no. 3, pp. 466-481, 2013.
- [6] A. Parsakhoo, M. A. Eshaghi and S. S. Joybari, "Design and evaluation of helicopter landing variants for firefighting in Golestan National Park, Northeast of Iran," *Caspian Journal of Environmental Sciences*, vol. 14, no. 4, pp. 321-329, 2016.
- [7] "The Directorate," The Civil Aviation Directorate of the Republic of Serbia, [Online]. Available: <https://cad.gov.rs/strana/16451/O-директоратy>. [Accessed 1 3 2025].
- [8] "The Agency," European Union Aviation Safety Agency, [Online]. Available: <https://www.easa.europa.eu/en/the-agency/the-agency>. [Accessed 1 3 2025].
- [9] "List of heliport operators with Agreement to operate an Heliport," The Civil Aviation Directorate of the Republic of Serbia, [Online]. Available: [https://cad.gov.rs/upload/aerodromi/2024/Evidencija%20aerodroma%20i%20helodroma/Evidencija%20heli-droma%20sa%20saglasno%C5%A1%C4%87u%20za%20kori%C5%A1%C4%87enje%20\(Izmena%2015\)%2001.11.2024.%20\(1\).pdf](https://cad.gov.rs/upload/aerodromi/2024/Evidencija%20aerodroma%20i%20helodroma/Evidencija%20heli-droma%20sa%20saglasno%C5%A1%C4%87u%20za%20kori%C5%A1%C4%87enje%20(Izmena%2015)%2001.11.2024.%20(1).pdf). [Accessed 1 3 2025].



- [10] "Traffic," [Online]. Available: www.mecavnik.html. [Accessed 1 3 2025].
- [11] "Field Listing – Heliports," The World Factbook, [Online]. Available: <https://www.cia.gov/the-world-factbook/field/heliports/>. [Accessed 1 3 2025].
- [12] N. Šarančić, "Search and Rescue Plan," 22 12 2020. [Online]. Available: <https://cad.gov.rs/upload/traganje/2021/SAR%20Plan%20u%20civilnom%20vazduhoplovstvu%20RS.pdf>. [Accessed 1 3 2025].
- [13] Official Gazette of the Republic of Serbia number 93/15, "Regulation on the manner of engagement of participants in aircraft search and rescue in civil aviation," 13 11 2015. [Online]. Available: http://demo.paragraf.rs/demo/combined/Old/t/t2015_11/t11_0115.htm. [Accessed 1 3 2025].
- [14] Official Gazette of the Republic of Serbia number 103/18, „Rules on the Conditions and Procedure for Issuing a Permit to Use a Heliport," 26 12 2018. [На мрежи]. Available: <https://cad.gov.rs/upload/regulativa/2018/1.%20Pravilnik%20o%20uslovima%20i%20postupku%20za%20izdavanje%20dozvole%20za%20koriscenje%20helidroma.pdf>. [Последњи приступ 1 3 2025].
- [15] S. K. Kumar, "Understanding helicopter performance," Vertical, 14 7 2020. [Online]. Available: <https://verticalmag.com/features/understanding-helicopter-performance/>. [Accessed 1 3 2025].



THE EFFECTIVENESS OF ONLINE EXECUTIVE COACHING IN LEADERSHIP DEVELOPMENT: A CASE STUDY

Maja Relić *,
[0009-0004-3622-8792]

Lazar Dražeta,
[0000-0001-5246-6468]

Borivoje Đokić
[0009-0007-6005-9556]

Singidunum University,
Belgrade, Serbia

Abstract:

This study explores the impact of one-on-one online executive coaching on developing active listening skills and leadership effectiveness. Given the increasing complexity of leadership roles in the VUCA (Volatile, Uncertain, Complex, and Ambiguous) business environment, engaging in effective communication and active listening has become a fundamental skill for executives. The research is based on a four-month coaching intervention with a Chief Executive Officer (CEO) in the energy infrastructure sector, utilizing structured exercises and assessments to track progress.

A two-category framework was developed to evaluate the impact of non-verbal communication and verbal communication traits on leadership effectiveness. The results indicate significant improvements across all dimensions of active listening, with particularly strong development in the category of verbal communication.

Moreover, the study underscores the effectiveness of online coaching, facilitated through Microsoft Teams, in delivering comparable, if not superior, results to traditional in-person coaching. The findings align with existing research demonstrating that virtual coaching enhances accessibility, flexibility, and scalability without compromising impact.

Overall, this study confirms that online executive coaching is an essential component of leadership development strategies, fostering stronger team engagement, problem-solving capabilities, and organizational effectiveness.

Keywords:

Executive Coaching, Online Coaching, Active Listening, Leadership Development, Microsoft Teams.

INTRODUCTION

Individuals continuously advance in their professional careers through role transitions, promotions, or shifts across different sectors. Each new position necessitates the acquisition of a diverse set of competencies that must be effectively developed [1]. Moreover, while formal education provides the foundational qualifications required for employment, it does not comprehensively equip individuals with the specialized skills essential for various professional roles. The concept of lifelong learning [2] has become increasingly relevant in contemporary professional environments, reflecting the necessity for continuous knowledge acquisition and personal development to meet the evolving demands of the modern workforce.

Correspondence:

Maja Relić

e-mail:

mrelic@singidunum.ac.rs





Education plays a crucial role in shaping both individuals and society, facilitating the acquisition of knowledge, skills, values, and attitudes through a complex web of interactions that fosters personal growth and active participation in the community [3]. In recent years, one area of professional education that has gained considerable attention is coaching [4]. Coaching can be described as a systematic process designed to assist individuals in achieving their goals, improving skills, and unlocking their potential through reflective dialogue and structured guidance [5]. This method serves as a valuable addition to traditional education, providing personalized opportunities for both personal and professional development, tailored to individual needs and aspirations [6]. It is most commonly applied in organizational contexts, with two main goals: (a) improving job performance and (b) offering career development support [7].

Coaching has become increasingly widespread and is now accessible not only to senior executives but also to a broader range of employees within organizations. In many cases, individuals may benefit from discussions with colleagues, friends, or mentors, as these interactions can provide new perspectives, facilitate problem-solving, and boost confidence in taking proactive steps toward personal or professional goals. Such engagements can be regarded as informal coaching sessions, characterized by their straightforward nature, which involves active listening, asking insightful questions, and refraining from offering prescriptive advice on resolving specific challenges.

Despite appearing straightforward, effective coaching is a structured discipline requiring formal training. While providing advice is a natural human tendency, professional coaching demands rigorous education and certification from reputable institutions. With the exponential growth of the coaching profession [4], the quality of training and accreditation has become crucial. Leading international organizations, such as the International Coach Federation (ICF), the Professional Coaches and Mentors Association (PCMA), and the Worldwide Association for Business Coaches (WABC), set the standard for coaching excellence, offering certification programs designed to ensure competency and ethical practice [8].

1.1. THE EVOLUTION OF EXECUTIVE AND ONLINE COACHING

Among the earliest and most established coaching models is executive coaching, which primarily targets senior leaders tasked with high-stakes decision-making. Executive coaching has gained widespread acceptance in corporate environments due to its potential to increase productivity by over 40% and generate a median return on investment (ROI) of 788% [9]. Traditionally, coaching sessions were conducted onsite and face-to-face, emphasizing direct interpersonal engagement. However, technological advancements and the growing need for flexibility have led to the increasing adoption of online coaching, a modality that offers equivalent, if not greater, effectiveness compared to in-person methods [10].

Empirical studies suggest that online coaching delivers outcomes comparable to traditional face-to-face coaching. A meta-analysis conducted by Theeboom et al. [11] demonstrated that virtual coaching interventions significantly enhance goal attainment, self-efficacy, and workplace performance, with no statistically significant difference between online and in-person formats. Additionally, Lindner et al. [10] found that online coaching fosters behavioral change and cognitive restructuring in ways that align with or even exceed the efficacy of traditional coaching models.

1.2. MICROSOFT TEAMS AS A PLATFORM FOR ONLINE COACHING

The integration of digital platforms into coaching practices has expanded accessibility, improved scheduling flexibility, and enabled seamless documentation of progress. Microsoft Teams is one of the most widely utilized platforms for virtual communication, offering a comprehensive suite of features tailored for professional development. The platform facilitates high-quality video conferencing, real-time chat, document sharing, and collaborative tools, creating an immersive coaching environment [12].

Key advantages of Microsoft Teams for online coaching include:

- session recording and note-sharing that enables the coachee to revisit discussions and track progress;
- breakout rooms, which allow for individualized coaching within group settings;
- Integration with productivity tools that support goal-setting through OneNote, Planner, and task management features;



- Cross-device accessibility that ensures uninterrupted participation from multiple locations.

In this case, the adoption of online coaching via the Microsoft Teams platform has transformed our leadership development intervention, making the coaching process more scalable, cost-effective, and easily affordable for both parties (i.e., coach and coachee).

1.3. LEADERSHIP, EMOTIONAL INTELLIGENCE, AND ACTIVE LISTENING

Leadership is the ability to influence teams toward shared objectives, requiring a combination of visionary thinking, emotional intelligence, and decision-making skills. According to Northouse [13], leadership is defined as a process by which an individual influences a group of individuals to achieve a common goal, emphasizing the interplay between influence, goal setting, and team dynamics. Contemporary leadership paradigms emphasize emotional intelligence (EI) as a crucial determinant of leaders' effectiveness. Goleman [14] claims that leaders with high EI demonstrate superior stress management, decision-making, and team cohesion, underlining the necessity of active listening as an EI-enhancing competency.

Active listening enables leaders to interpret verbal and nonverbal cues, fostering empathy as one of the core dimensions of emotional intelligence. Research highlights that executives who undergo structured coaching programs incorporating active listening techniques report measurable improvements in team engagement, collaboration, and workplace morale [15]. Online coaching platforms further augment this development by facilitating interactive exercises, feedback loops, and asynchronous learning opportunities.

This paper aims to provide an in-depth analysis of active listening as a critical leadership skill, examined through the lens of coaching as a leading professional education method. Additionally, it will explore the role of digital coaching solutions in enhancing leadership effectiveness, particularly in the context of an increasingly virtual and globalized business environment.

2 METHODOLOGY

This study focuses on the Chief Executive Officer (CEO) of an organization that plays a significant role in the energy infrastructure sector. As a condition for publishing this paper, both the company and its CEO have been kept confidential. The company employs around 1,000 people and was established with the primary goal of meeting the needs of the energy industry, particularly in the areas of energy infrastructure development and maintenance.

The focal point of this study is the CEO (the coachee), a middle-aged man who has been with the company for 16 years and manages a team of 17 direct reports, predominantly consisting of directors. In 2023, the CEO initiated efforts to improve the development of his leadership team by conducting a 360-degree feedback survey through an external firm. The survey results, based on feedback from his direct reports, revealed concerns about his communication style, which was perceived as significantly inadequate. In response to this feedback, the CEO decided to work with a professional coach to facilitate the necessary improvements. Following an initial one-on-one session, it was determined that improving the CEO's active listening abilities would be the most effective way to enhance his communication and, consequently, his leadership.

During the initial coaching sessions, an assessment of the current situation was made, and goals for the future were established. Due to the coachee's frequent business trips, the sessions were conducted remotely every week, with occasional in-person meetings, over four months. The coach utilized the widely adopted GROW model, which includes four distinct stages that guide the coaching process [16, 17].

- *G (Goal)*. The main objective was to enhance the coachee's communication by focusing on the improvement of active listening skills.
- *R (Reality)*. To evaluate the current level of coachee competence and subsequent improvements resulting from the coaching process, 9 key criteria were established through a questionnaire designed to appraise active listening skills. The coachee was observed and evaluated in various regular situations while leading discussions with direct reports, using the rating format of the Mixed Standard Scale (MSS): below standard (score "1" in table), at standard (score "2"), and above standard (score "3") [18].



- *O (Option)*. A range of practical exercises was identified to strengthen the coachee's active listening skills, which were then incorporated into their everyday activities. Detailed descriptions of these exercises are provided later in the paper.
- *W (Will)*. The coachee's drive to progress was sustained through clear and actionable steps. It was essential for the coachee to consistently follow through with the exercises provided by the coach. Consequently, most coaching sessions were held online.

To evaluate the improvements in active listening skills throughout the process, the Harvard self-assessment questionnaire was used. This questionnaire consists of 13 questions that evaluate active listening skills. Although it is a self-assessment tool, the coach used it to evaluate the coachee's improvements at the beginning and end of the coaching process, based on a simple evaluation of whether certain behaviors are demonstrated always, sometimes, or rarely [19].

The nature of the criteria evaluation and sample size directed the methodology used for data analysis. The statistical software used in this research was SPSS version 29.0 [20]. A set of 9 evaluation criteria was further arranged into two distinct categories, namely non-verbal and verbal communication. Statistical hypothesis testing is an essential tool for inferential statistics that enables researchers to make informed decisions about population parameters based on sample statistics. One common hypothesis test used in this study for comparing two sample means is the Two-Sample Z-test, which proves more precise than the T-test, providing that standard deviations are known from previous research. Dyrbye et al. [21] reported standard deviation of pre-coaching was between 0.29 (29%) to 0.33 (33%), while post-coaching standard deviation was between 0.33 (33%) and 0.34 (34%). As a final result of analysis researcher is presented with the *p-value* that shows the significance of the observed differences between variables, based on the statistical evidence.

The initial exercises were designed with a progressive approach, starting with simpler tasks that gradually increased in complexity to reinforce prior learning and skill development. One such exercise involved a memory challenge, where a sequence of 15 words was spoken aloud to the coachee, who was then required to remember and repeat them after a specified time interval. This task was specifically aimed at strengthening active listening, improving concentration, and enhancing memory retention, all of which are crucial for sustained attention in professional interactions.

To further reinforce these cognitive abilities, the coachee was assigned a meditation-based focus exercise as part of independent practice between coaching sessions [22]. This technique required directing attention to a single object for a short duration, with the initial goal of maintaining focus for one minute. Upon successful completion, the duration was incrementally extended by one minute each day. This gradual increase aimed to cultivate sustained attention, enhance cognitive control, and support overall mental resilience, which are essential attributes for effective leadership and communication.

The next exercise incorporated the use of narratives with illogical patterns and an absence of clear connections [23]. The coachee was exposed to three to four such narratives and was required to retain as many details as possible to respond accurately to a series of complex follow-up questions. This exercise was designed to challenge conventional cognitive processing by encouraging engagement with non-linear and abstract thinking patterns. Research suggests that children tend to approach problem-solving with a greater degree of creativity and open-mindedness, whereas adults, shaped by societal norms and cognitive conditioning, often rely on structured and analytical reasoning, which can constrain innovative capacity [24]. By participating in imaginative and unconventional exercises, the coachee was able to re-engage cognitive processes associated with creative problem-solving, resembling the flexible thinking exhibited in childhood.

To develop the ability to paraphrase and avoid interruptions, specific exercises were implemented. Tangram puzzles require the coachee to manipulate geometric shapes to construct a final image representing recognizable objects, such as a house or aircraft. Throughout this exercise, the coach offered spoken guidance, guiding the coachee in putting together the designated figure while holding the reference image. In an unguided image creation exercise, the coachee received a sequence of step-by-step verbal instructions to draw various shapes but without access to the final visual reference. This approach necessitated active engagement in seeking clarifications, accurately paraphrasing directives, and demonstrating patience while awaiting additional guidance.

All activities carried out were intentionally crafted to improve the coachee's attentiveness and mental agility, emphasizing the value of integrating creative tasks into executive coaching as a way of promoting leadership growth.



3 RESULTS AND DISCUSSION

As the process of coaching continued, early results started to emerge. The initial exercise, a memory game, showed significant advancements in cognitive retention. Initially, when the coachee was presented with a set of 15 randomly selected words with no logical associations, he was able to recall only 6 to 7 terms. However, with continued practice, his performance significantly improved, ultimately reaching 12 to 13 accurately recalled words. These findings align with research indicating that structured memory training enhances working memory capacity and cognitive flexibility [25].

Daily meditation practices also played a significant role in achieving positive results. With each session, the meditation duration was progressively increased by 1 to 2 minutes, leading to enhanced concentration and sustained attention. Prior studies have shown that mindfulness-based meditation strengthens attentional control and improves cognitive performance, supporting the observed improvements in the coachee's ability to maintain focus during tasks [26].

This subsequently increased cognitive focus, which resulted in measurable progress in the illogical stories exercise. Initially, the coachee provided only 1 to 2 correct answers, but after several weeks of consistent practice, his accuracy increased to 6 to 7 correct responses. This progress can be attributed to the coachee's improved capacity to transcend traditional logical limitations, thereby activating the right hemisphere of the brain, which is linked to creativity and intuitive thinking. Research confirms that engaging in tasks that require unconventional problem-solving activates right-hemispheric brain functions, fostering greater cognitive flexibility and creative thinking [27].

These findings underscore the effectiveness of integrating memory training, meditation, and creative problem-solving exercises into executive coaching, facilitating both cognitive and leadership development.

After four months of continuous one-on-one online coaching, a comprehensive assessment was conducted to evaluate improvements in the coachee's workplace communication with direct reports, external clients, and partners. The results are illustrated in Table 1.

Table 1. Coaching results for active listening skills

Category	Evaluation criteria	Before the coaching process	After the coaching process	Description of improvements
Non-verbal communication	Eye contact	2	3	Longer and more confident eye contact, avoiding gaze aversion that enhances connection and credibility in communication.
	Body language	2	3	Use open and relaxed gestures, avoid closed posture cues, thereby appearing more approachable and engaged.
	Non-verbal cues	1	2	Improved recognition of non-verbals from interlocutors and adjusts facial expressions and movements to convey understanding and empathy.
	Demonstrating interest	1	2	Actively nods, uses affirming expressions such as "I understand" and "Yes, I agree," and maintains an engaged facial expression.
Verbal communication	Seeking clarification	1	2	Asks additional questions when information is unclear and verifies understanding before responding or concluding.
	Relevant Questions	1	3	Asks clear, open-ended questions that encourage the interlocutor to participate, in contrast to the previous period when all of his questions were closed-ended.
	Paraphrasing	1	2	Actively uses paraphrasing in conversation, framing sentences with "If I understand you correctly..." and conveying understanding in her own words.
	Avoiding interruptions	1	2	Consciously refrains from interrupting the speaker, patiently listens before responding, demonstrating greater respect for others' perspectives.
	Summarizing	2	2	Summarize the key points of the discussion, taking notes in a notebook.



By inspecting Table 1, significant progress in the development of the coachee's active listening skills can be observed after the completion of the coaching process. Two categories defined in our study (non-verbal and verbal communication) are statistically significantly different in favor of after-coaching. Verbal communication traits show a stronger increase of 83.33% (with statistical significance of $p=0,006$) versus non-verbal communication traits of 66.67% (with statistical significance of 0,015). This can be explained by the fact that exercises conducted through coaching were mainly aimed at improving verbal communication skills.

Beyond communication improvements, the coaching process created a ripple effect throughout the organization, yielding key leadership enhancements, observed by the organization's owners and other team members.

- *Building trust and engagement.* Strengthened non-verbal communication helped foster a more transparent and open leadership style, increasing team morale.
- *Problem-solving facilitation.* The shift towards coaching-based inquiry enabled deeper discussions, encouraging proactive decision-making.
- *Delegation and accountability.* A more structured communication approach clarified expectations, empowering employees and reducing operational inefficiencies.
- *Leadership modeling.* By demonstrating active listening and structured communication, the coachee set a new behavioral standard, reinforcing a culture of collaboration.

The four-month online coaching intervention resulted in substantial improvements in the coachee's communication effectiveness, particularly in verbal engagement. These enhancements directly influenced leadership performance, fostering a more collaborative, transparent, and high-performing work environment. The findings underscore the critical role of structured online coaching in leadership development, providing a scalable and effective approach to improving executive communication skills.

4 CONCLUSION

This research highlights the essential impact of professional education techniques, especially executive coaching, on boosting leadership effectiveness. The deliberate cultivation of active listening skills leads to enhanced communication, stronger interpersonal relationships, and improved team dynamics—key elements of effective leadership [13], [14]. By incorporating coaching strategies into professional development, organizations can provide leaders with the skills needed to manage complex interpersonal dynamics and promote long-term team success.

In today's VUCA (volatile, uncertain, complex, and ambiguous) business world, the capacity for effective communication through active listening is becoming more crucial—not just for individual leadership growth but also for driving overall team success. As the idea of continuous learning gains momentum, professional development methods like coaching have proven essential for shaping adaptive and impactful leaders in the fast-evolving workplace [10].

The evolution of digital communication tools has contributed to the growing acceptance and effectiveness of online coaching. Recent empirical studies indicate that virtual coaching yields comparable and, in some cases, even superior results to traditional in-person coaching, particularly in terms of accessibility, scheduling flexibility, and scalability [4], [11]. The integration of digital platforms, such as Microsoft Teams, has facilitated seamless interactions between coaches and coachees, ensuring continuity in leadership development regardless of geographical constraints [12].

This study confirms that executive coaching, both online and in-person, has become an indispensable component of leadership development strategies. As organizations continue to embrace digital transformation, online coaching is positioned to play an increasingly prominent role in shaping future leaders. The findings reinforce the idea that effective coaching is not determined by the medium through which it is delivered but by the quality of interaction, structured methodology, and the commitment of both coach and coachee.



REFERENCES

- [1] R. Charan, S. Drotter, and J. Noel, *The Leadership Pipeline: How to Build the Leadership-Powered Company*, Jossey-Bass, John Wiley & Sons, 2011.
- [2] European Commission, *The European Skills Agenda for Sustainable Competitiveness, Social Fairness and Resilience*, in *Proceedings of the European Commission*, 2007.
- [3] G. Biesta, *Learning about Learning: A Philosophy of Education*, in *Proceedings of Routledge*, London, 2011.
- [4] International Coach Federation, *ICF Global Coaching Study: Executive Summary*, commissioned by ICF and conducted by PricewaterhouseCoopers (PwC), 2023. [Online]. Available: https://coachingfederation.org/app/uploads/2023/04/2023ICFGlobalCoachingStudy_ExecutiveSummary.pdf
- [5] J. Whitmore, *Coaching for Performance: The Principles and Practice of Coaching and Leadership*. London, U.K.: Nicholas Brealey Publishing, 2017.
- [6] A. M. Grant, "The Impact of Coaching on Development," in *The Handbook of Coaching Psychology*. New York, NY, USA: Routledge, 2017.
- [7] State Services Commission, *Coaching: Guidance for Managers and Their Staff on Using Coaching for Development and Performance*, State Services Commission (Te Komihana O Ngā Tari Kāwanatanga), pp. 1–33, 2003.
- [8] M. Homan and L. J. Miller, *Coaching in Organizations – Best Coaching Practices from The Ken Blanchard Companies*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2008.
- [9] J. Burke, "Seven Reasons Growth Companies Need Corporate Executive Coaching," Creo Consulting, 2024. [Online]. Available: <https://creoconsulting.com/seven-reasons-growth-companies-need-executive-coaching/>.
- [10] H. Lindner, K. L. Peyer, Y. Bai, L. D. Ellingson, and G. J. Welk, "Comparison of In-Person and Online Motivational Interviewing-Based Health Coaching," *Journal of Medical Internet Research*, vol. 19, no. 10, p. e358, 2017. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/29212349/>.
- [11] T. Theeboom, B. Beersma, and A. E. M. van Vianen, "Does coaching work? A meta-analysis on the effects of coaching on individual-level outcomes in an organizational context," *The Journal of Positive Psychology*, vol. 9, no. 1, pp. 1–18, 2014. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/17439760.2013.837499>.
- [12] Microsoft, "Microsoft Teams – The digital hub for teamwork in Microsoft 365," 2023. [Online]. Available: <https://www.microsoft.com/en-us/microsoft-teams/>. [Accessed: Mar. 21, 2025].
- [13] P. G. Northouse, *Leadership: Theory and Practice*, 8th ed. Thousand Oaks, CA, USA: Sage Publications, 2018.
- [14] D. Goleman, *Emotional Intelligence*, 10th Anniversary Ed. New York, NY, USA: Bantam Books, 2005.
- [15] J. W. Neal and R. L. Edwards, "Strategic Listening: How Managers, Coworkers, and Organizations Can Become Better at Listening," *Journal of Organizational Communication*, vol. 45, no. 3, pp. 215–230, 2023. [Online]. Available: https://www.researchgate.net/publication/374565789_Strategic_Listening_How_Managers_Coworkers_and_Organizations_Can_Become_Better_at_Listening. [Accessed: Mar. 21, 2025]
- [16] J. Whitmore, *Coaching for Performance: GROWing Human Potential and Purpose - The Principles and Practice of Coaching and Leadership*. London, U.K.: Nicholas Brealey Publishing, 2009.
- [17] M. Turner, "The GROW Model," *Mentoring for Change*, 2009. [Online]. Available: www.mentoringforchange.co.uk/classic/index.php. [Accessed: Mar. 21, 2025]
- [18] F. Blanz and E. E. Ghiselli, "The Mixed Standard Scale: A New Rating System," *Personnel Psychology*, 2006. [Online]. Available: https://www.researchgate.net/publication/229622618_The_Mixed_Standard_Scale_A_New_Rating_System. [Accessed: 28-Mar-2025]
- [19] Harvard Business School Publishing, "Active Listening Self-Assessment," *Scribd*, [Online]. Available: <https://www.scribd.com/doc/254235970/Active-Listening-Self-Assessment>. [Accessed: 28-Mar-2025]
- [20] IBM, *IBM SPSS Statistics 29.0*. [Online]. Available: <https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-29>. [Accessed: Mar. 31, 2025]
- [21] L. M. Dyrbye, S. E. West, M. S. Sinsky, et al., "Effect of a professional coaching intervention on the well-being and distress of physicians: A pilot randomized clinical trial," *JAMA Internal Medicine*, vol. 179, no. 10, pp. 1406–1414, Oct. 2019. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/31380892/>. [Accessed: Mar. 31, 2025]
- [22] I. Pavlović and M. Šimić, "The effects of mindfulness practice on attention: A review of research," *Psychological Topics*, vol. 28, no. 3, pp. 451–472, 2019. [Online]. Available: <https://hrcak.srce.hr/file/339285>. [Accessed: Mar. 21, 2025]
- [23] R. Rajović, "NTC quiz," YouTube, 2022. [Online]. Available: <https://www.youtube.com/watch?v=uBYs30IJpgA>. [Accessed: Mar. 21, 2025]
- [24] K. Robinson, "Do Schools Kill Creativity?" *TED Talk*, 2006. [Online]. Available: https://www.ted.com/talks/sir_ken_robinson_do_schools_kill_creativity?subtitle=en. [Accessed: Mar. 21, 2025]



- [25] S. M. Jaeggi, M. Buschkuhl, J. Jonides, and W. J. Perrig, "Improving Fluid Intelligence With Training on Working Memory," *Proceedings of the National Academy of Sciences*, vol. 105, no. 19, pp. 6829–6833, 2008. [Online]. Available: <https://doi.org/10.1073/pnas.0801268105>
- [26] L. F. Barrett and T. Wager, "The Effects of Mindfulness Meditation on Attention and Cognitive Flexibility," *Journal of Cognitive Neuroscience*, vol. 29, no. 8, pp. 1456–1468, 2017. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/>
- [27] M. Beeman and J. Bowden, "The Right Hemisphere's Role in Creative Problem Solving," *Neuropsychologia*, vol. 46, no. 1, pp. 281–287, 2008. [Online]. Available: <https://doi.org/10.1016/j.neuropsychologia.2007.07.012>



ETHICAL CONCERNS AND MASS AGENTIC AI ADOPTION

Sonja Jovanović^{1*},
[0009-0009-3538-8096]

Lazar Dražeta²,
[0000-0001-5246-6468]

Aleksandar Petrović²,
[0000-0003-3324-3909]

Nebojša Bačanin Džakula²,
[0000-0002-2062-924X]

Miodrag Živković²
[0000-0002-4351-068X]

¹Ernst & Young,
Belgrade, Serbia

²Singidunum University,
Belgrade, Serbia

Abstract:

The arrival of agentic artificial intelligence (AI) marks a significant jump in the evolution of intelligent systems. Unlike AI so far, which primarily operates within predefined parameters, agentic AI possesses the capability to autonomously perceive, reason, act, and learn from its environment. This new paradigm enables AI agents to solve complex, multi-step problems and interact with various tools and data sources independently. Rather than executing commands or routines set by humans, Agentic AI systems are intended to adjust their strategies and explore their environments as needed. Agentic AI's sophisticated reasoning and iterative planning abilities are settled to transform numerous industries. From optimizing supply chains and personalizing customer service to enhancing healthcare and automating complex tasks, the potential applications are vast and varied. However, with these advancements come challenges, including the need to address biases, ensure ethical use, and establish robust regulatory frameworks. This paper explores the foundational principles of Agentic AI, its current and potential applications, and the ethical considerations that must guide its development and deployment. By examining these aspects, the aim is to provide a good understanding of how Agentic AI can be used to drive innovation while mitigating associated risks.

Keywords:

Agentic AI, Evolution, Autonomous Decision Making, Use Cases.

INTRODUCTION

Agentic artificial intelligence (AI) is an autonomous or semi-autonomous software entity that can process data, form decisions and take actions in the digital and physical world. It can process information and perform tasks, rather than just generating answers like many of the chatbots seen in recent years.

AI chatbots use generative AI to provide responses. It is important to highlight that chatbot response will be original, but highly dependent on already known or most frequently repeated information. When a person asks a question, the chatbot utilizes natural language processing to respond. Agentic AI goes beyond [1]. It stands out due to its autonomy and ability to independently pursue specific goals while learning from feedback. Agentic AI continuously monitors and processes data inputs, as well as inputs from the users.

Correspondence:

Sonja Jovanović

e-mail:

sonja.jovanovic@rs.ey.com





Based on those inputs, it autonomously makes informed decisions and takes actions, if needed. It also learns from its experiences and mistakes. In essence, Agentic AI can autonomously perform complex tasks in real time [2]. The adoption of generative AI has surpassed all the entertainment platforms in terms of adoption time, as illustrated in Figure 1. The speed at which it was adopted is immense considering that the entertainment industry is of much higher interest to the general public. However, the impact of AI and visions of a cyberpunk future have prevailed against the interests of the average person.

When examining the differences between generative AI and agentic AI, three main points are notable: it focuses on decision-making rather than content creation, it operates independently, targeting specific goals such as increasing sales, enhancing customer satisfaction, or improving employee experiences, without needing human prompts, and agentic AI can perform complex tasks and independently search through various databases and data sources.

Mass application of agentic AI is the next step, as its use cases are vast. This work reviews its numerous applications and provides a perspective on the ethical dilemmas associated with it. Without considering the possible job shortages that such systems would produce, healthcare concerns are first to be raised [4]. There must exist a standard before approving that the AI in such an environment is safe for human use. Therefore the goal of this work is to provide a connection between the use cases and risks that they bring with their use.

2. AGENTIC AI USE CASES

Agentic AI can boost productivity and efficiency for security, financial operations, software development, customer engagement, digital marketing, knowledge acquisition, and more [5]. As Agentic AI can self-correct and improve over time, it is expected to accelerate, optimise and boost productivity and system performance.

Predictions from Oracle and Gartner are saying that Agentic AI will soon find its way into 33% of enterprise software applications, replacing 20% of human interactions and making 15% of day-to-day workplace decisions by 2028 [6]. In addition, the global market for Agentic AI is projected to reach \$120 billion by 2030, driven by ever-increasing investments in AI technologies and their applications across various industries [7].

Autonomous vehicles are well-known examples of agentic AI today. These vehicles use multimodal models to process data from sensors like cameras, light detection and ranging (LiDAR), a technology that measures distance with laser beams, and radar. This helps them understand their surroundings, recognise objects, and make decisions for safe navigation. They use predictive algorithms to respond faster than humans, preventing dangerous situations. By processing sensor data and using deep learning algorithms, autonomous vehicles can determine the best route, reduce travel time, and identify obstacles in real time. Waymo's driverless cars, originating from Google's self-driving car project is the world's earliest autonomous ride-hailing service and a great example of Agentic AI in action [8].

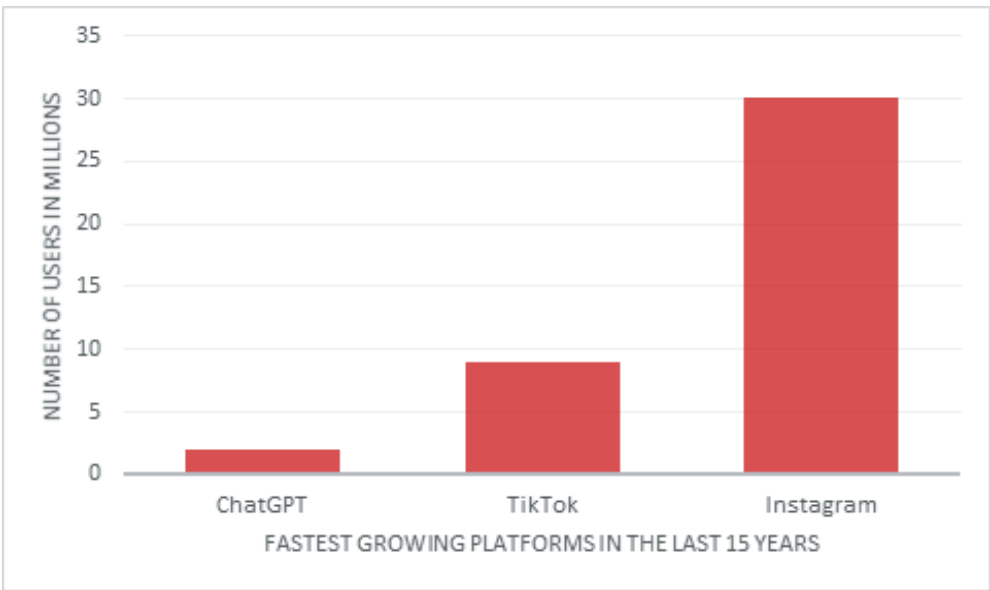


Figure 1. Speed of adoption of revolutionary technological platforms [3]



Agentic AI systems, with their advanced decision-making and execution abilities, are excellent platforms for experimentation and innovation. Multi-agent AI models can quickly scan and analyse large research spaces, like scientific articles and databases, much faster than human researchers. SciAgents, developed by MIT, are robot scientists to create research plans and a critic agent to review and improve them [9]. As an example, these two AI agents identified a new biomaterial made from silk and dandelion pigments, which has better mechanical and optical properties and requires less energy to produce.

An Agentic customer service agent can predict if a delivery will be late, notify the customer about the delay, and offer a discount to ease their disappointment. Agentic AI can search through thousands of databases and apps to solve customer queries and complaints [10]. These chatbots learn from each interaction and suggest actions for human agents. They can also check its content for accuracy and compliance and recommend improvements to the customer knowledge base.

For sales agents, the essential task of identifying and nurturing sales leads can often be overwhelmed by numerous emails, paperwork, and other routine but necessary administrative duties. AI systems designed for sales can significantly reduce the time spent on these activities. For instance, Salesforce has introduced its Agent Force service development rep [11] to support human sales teams. Using large language models (LLM), this AI can understand customer messages, suggest follow-up actions, schedule meetings, answer inquiries, and create responses that match the company's brand voice. Additionally, the agent force sales coach offers personalized feedback to sales agents and provides learning opportunities through virtual role-play sessions.

The ability of agentic AI systems to adapt to various environments, understand human emotions, and demonstrate empathy makes them well-suited for non-routine, soft-skills tasks in fields like healthcare and caregiving. Hippocratic AI [12], a healthcare company based in California, has developed a range of AI agents specialized in different aspects of healthcare and social support. Among these agents is Sarah – who is known for her warmth and understanding in assisting with daily living activities. Sarah can engage with patients about their day, manage meal plans and transportation, and remind them to take their medication. Another agent, Judy, supports patients with pre-operative procedures by providing reminders about arrival times, locations, and instructions on fasting or medication adjustments.

In the gaming industry, Agentic AI has a lot of things to offer. Enhancing the behaviour of bots in video games is one of them. Non-player characters (NPC) are characters in video games that are not controlled by the human player – like enemies, allies, merchants, quest givers, etc. While traditional NPCs follow pre-programmed scripts or simple learning algorithms, agentic AI enables NPCs to exhibit more complex and adaptive behaviours. By using the reinforcement learning technique, NPCs can learn from their environment and improve their actions over time, based on rewards and penalties [13]. For example, NPCs can engage in more natural and varied conversations, responding to the player's choices, or if a human player acts aggressively, NPCs might become more hostile or wary.

There is significant potential for using Agentic AI in human resource management. For example, in talent management, AI agents can analyse employee performance data, identify skill gaps, and recommend personalized development plans [14]. This approach ensures that employees receive targeted training and opportunities for career growth. Another valuable application is in employee engagement and retention. Agentic AI can monitor employee satisfaction through sentiment analysis of feedback and interactions. It can also suggest interventions to enhance employee experience and engagement, that result in reducing turnover.

The adoption of AI in some countries is high according to the IBM AI adoption index [15]. This index shows the percentage of enterprises that actively deploy technologies reliant on AI. The data is self-reported by the companies included in the survey. The country with the highest AI adoption index is India, amounting to almost 60% usage. The country with the lowest adoption rate was Australia, however, this should not be considered generally as bad, since most of the countries do not even make it on this list.

There are a lot of talks about AI, but what will really separate companies from their competitors, when it comes to AI agenda – is reliable, legal, and proprietary data at scale. According to EY research, 83% of senior leaders recognise there is a gap in their capabilities and believe that their AI adoption would accelerate if they had stronger data infrastructure. Furthermore, 67% of senior leaders also admit their lack of infrastructure is actively holding back AI adoption [16].

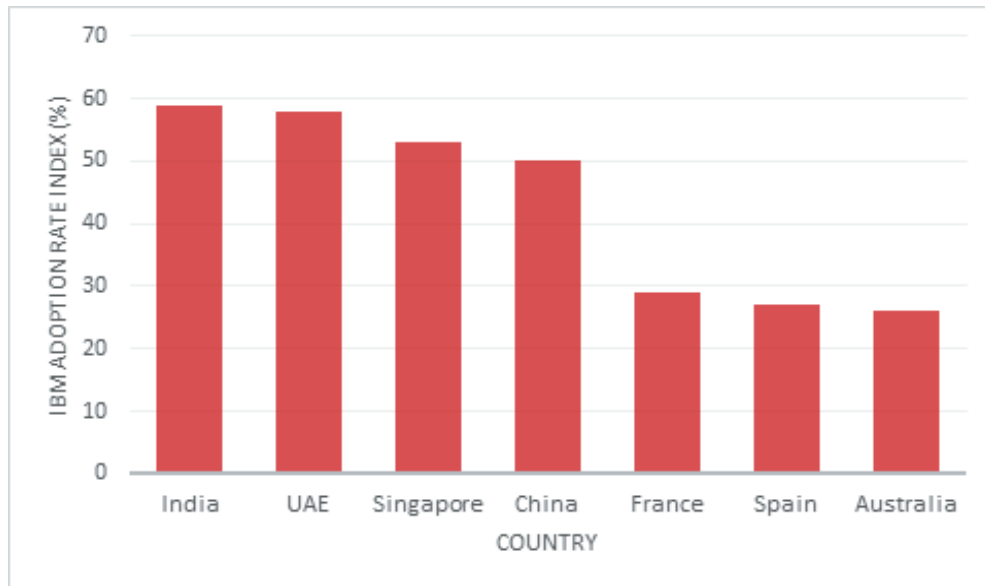


Figure 2. IBM AI adoption rate index [15]

This concept emphasizes that the quality of output from an AI system is directly dependent on the quality of the input data it receives. For Agentic AI, which relies on vast amounts of data to learn, reason, and make decisions autonomously, ensuring high-quality, accurate, and unbiased input data is crucial. Poor quality data can lead to flawed reasoning, biased outcomes, and ineffective actions, undermining the potential benefits of such advanced AI systems [17]. For this reason, Agentic AI may not be perceived as completely trustworthy. Similar to how human teams can struggle with unclear or poorly communicated goals, Agentic AI systems can also fail if their objectives are not clearly defined.

3. RISK MANAGEMENT

As AI models become more advanced and agentic, they may also become less predictable and sometimes more ethically ambiguous. To address these risks, organizations should establish guidelines, risk management frameworks, and governance protocols. These measures will help ensure that Agentic systems prioritise legality, data privacy, security, human rights, and corporate accountability.

It is crucial to develop strategies for securely integrating this technology with existing system architecture. Additionally, it is important to create fail-safe mechanisms to prevent systemic failures. Regular behavioural monitoring, observation, and system analysis should also be conducted.

The most important feature is the presence of a human-in-the-loop (HITL), ensuring that employees are strategically positioned to monitor and oversee agentic operations. Information security teams considering AI agent integration should review the entire workflow from an automation-only perspective and then determine how to reintegrate human personnel into the agentic workflow. This requires education and training to ensure the workforce understands AI capabilities, benefits, and limitations.

While the risks and vulnerabilities are addressed, there are also costs associated with deploying and securing AI agents in enterprise environments. Additionally, there are unique governance challenges related to security assurance, regulation, legal accountability, data equity, and interoperability.

With great power comes great responsibility. As AI initiatives yield benefits, senior leaders must address the ethical implications and risks. Interest in responsible AI among senior leaders has increased over the past year, 61% vs. 53% six months ago [16]. Respondents also indicate that this interest will continue to grow over the next year. Consequently, more organizations are dedicating time to training employees in responsible AI and increasing transparency with customers about AI usage.

The variety of fields affected by AI also poses another threat to safety. This technology is making its way into critical fields like EEG predictions [18], intrusion detection [19], crop yield [20], gold price prediction [21], and some not-so-critical but nonetheless very important like software defect prediction [22], fake news detection [23], and machine learning optimization [24].

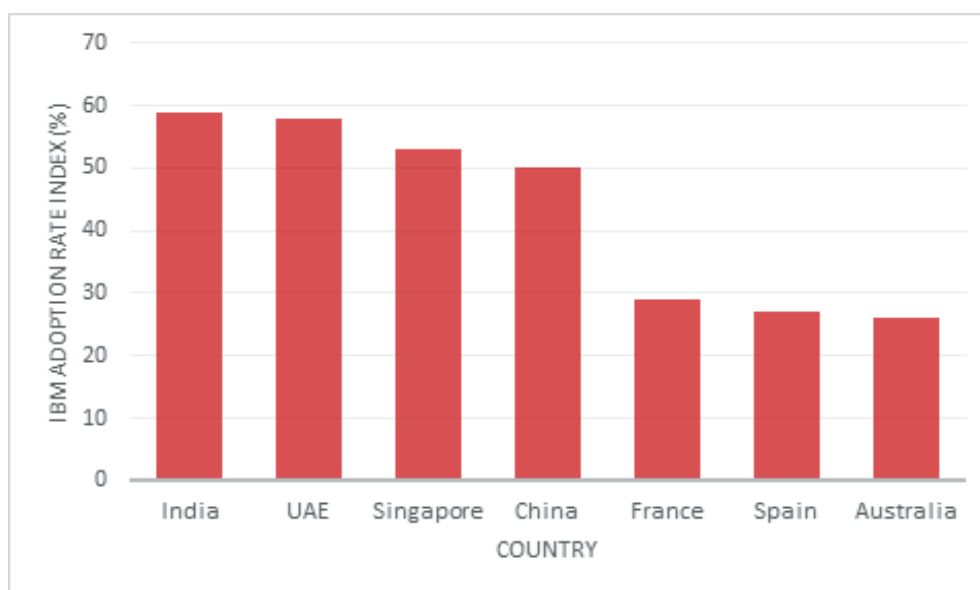


Figure 3. Hallucination rates of the popular generative agents [25]

The risks of AI use are depicted in Figure 3. It is shown how much of the data was hallucinated by generative agents. This is one of the most important topics regarding generative agents. The phenomenon of hallucination in generative agents can be described as making up information. This is a tendency of such agents when there is no usable data and no information in their knowledge base. The importance of this phenomenon is for the real-world use of such agents. They cannot be applied in critical fields until the rate of hallucinations is brought down to zero.

4. CONCLUSION

Agentic AI represents a significant advancement in AI, with the potential to transform various sectors. These systems, characterized by their ability to adapt, reason, and exhibit empathy, offer promising solutions for non-routine, soft-skills tasks. Agentic AI systems are still in the early stages of development. Although they have advanced reasoning and execution abilities, they do not eliminate traditional workforce management challenges – they transform them. Similar to managing human teams, managers must consider team composition and role selection and set appropriate goals to ensure the success of Agentic AI or hybrid teams. Additionally, managers need to determine when Agentic AI systems can be trusted to make decisions and when human intervention is necessary.

To utilise the full potential of Agentic AI, it is essential to establish comprehensive guidelines, risk management frameworks, and governance protocols. Ensuring the presence of a HITL is crucial for monitoring and overseeing AI operations. Upskilling, reskilling, education, and training programs are necessary to equip the workforce with a functional understanding of AI capabilities, benefits, and limitations. Further advancements in AI should bring full autonomy, which raises the question if it should be researched at all.

The lack of regulation is still a massive issue and the biggest threat from AI. This is an experimental technology and should be treated as such, which requires some security precautions. Current awareness on this topic is low, and that is why this paper is important as it serves to provide a review of the current state of agentic AI and the risks of its use.



REFERENCES

- [1] D. B. Acharya, K. Kuppan and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey," in *IEEE Access*, vol. 13, pp. 18912–18936, 2025, doi: 10.1109/ACCESS.2025.3532853.
- [2] B. L. Aylak, "SustAI–SCM: Intelligent Supply Chain Process Automation with Agentic AI for Sustainability and Cost Efficiency," *Sustainability*, vol. 17, no. 6, p. 2453, 2025.
- [3] K. Hu, "ChatGPT sets record for fastest-growing user base – analyst note," Reuters, 2023 February 2023. [Online]. Available: <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>. [Accessed 15 March 2025].
- [4] T. A. Bach, J. K. Kristiansen, A. Babic and A. Jacovi, "Unpacking Human–AI Interaction in Safety–Critical Industries: A Systematic Literature Review," in *IEEE Access*, vol. 12, pp. 106385–106414, 2024, doi: 10.1109/ACCESS.2024.3437190.
- [5] S. Samdani, K. Paul, and F. Saldanha, "Agentic AI in the Age of Hyper–Automation," *World Journal of Advanced Engineering Technology and Sciences*, vol. 08, no. 01, pp. 416–427, 2023.
- [6] D. Sun, "Capitalize on the AI Agent Opportunity," Gartner, 27 February 2025. [Online]. Available: <https://www.gartner.com/en/articles/ai-agents>. [Accessed 10 March 2025].
- [7] L. Craig, "What is agentic AI? Complete guide," TechTarget, September 2024. [Online]. Available: <https://www.techtarget.com/searchenterpriseai/definition/agentic-AI>. [Accessed 10 March 2025].
- [8] B. Xia, J. Zhou, F. Kong, Y. You, J. Yang, and L. Lin, "Enhancing 3D object detection through multi-modal fusion for cooperative perception," *Alexandria Engineering Journal*, vol. 104, pp. 46–55, 2024.
- [9] A. Ghafarollahi, and M. J. Buehler, "SciAgents: Automating Scientific Discovery Through Bioinspired Multi–Agent Intelligent Graph Reasoning," *Advanced Materials*, p. 2413523, 2024.
- [10] D. Leocádio, L. Guedes, J. Oliveira, J. Reis, and N. Melão, "Customer service with AI–powered human–robot collaboration (HRC): A literature review," *Procedia Computer Science*, vol. 232, pp. 1222–1232, 2024.
- [11] Salesforce, "Meet Einstein SDR and Einstein Sales Coach: Two New Autonomous AI Sales Agents to Scale Your Sales Team," Salesforce, 22 August 2024. [Online]. Available: <https://www.salesforce.com/news/stories/einstein-sales-agents-announcement/>. [Accessed 10 March 2025].
- [12] H. AI, "Hippocratic AI: Safety Focused Generative AI for Healthcare," Hippocratic AI, 2025. [Online]. Available: <https://www.hippocraticai.com/>. [Accessed 10 March 2025].
- [13] A. Maciá–Lillo, A. Jimeno–Morenilla, H. Mora and E. Duta, "Hybrid Architecture for AI–Based RTS Games," in *IEEE Transactions on Games*, doi: 10.1109/TG.2025.3533949.
- [14] M. Rožman, D. Oreški, and P. Tominc, "Integrating artificial intelligence into a talent management model to increase the work engagement and performance of enterprises," *Frontiers in psychology*, vol. 13, p. 1014434, 2022.
- [15] E. Shein, "IBM: While Enterprise Adoption of Artificial Intelligence Increases, Barriers are Limiting Its Usage," TechRepublic, 12 January 2024. [Online]. Available: <https://www.techrepublic.com/article/ibm-global-ai-adoption-index/>. [Accessed 10 March 2025].
- [16] L. McWilliams, "EY research: Artificial intelligence investments set to remain strong in 2025, but senior leaders recognize emerging risks," EY, 10 Dec 2024. [Online]. Available: https://www.ey.com/en_us/newsroom/2024/12/ey-research-artificial-intelligence-investments-set-to-remain-strong-in-2025-but-senior-leaders-recognize-emerging-risks?utm_source=chatgpt.com. [Accessed 10 March 2025].
- [17] A. Majeed and S. O. Hwang, "When Poor–Quality Data Meet Anonymization Models: Threats and Countermeasures," in *IEEE Access*, vol. 13, pp. 49457–49475, 2025, doi: 10.1109/ACCESS.2025.3552412.
- [18] N. Bacanin, L. Jovanovic, A. Toskovic, M. Zivkovic, A. Petrovic, and M. Antonijevic, "Anomalous EEG Signal Time Series Classification Using Modified Metaheuristic Optimized RNN," in *International Conference on Communication and Intelligent Systems*, Springer, 2023, pp. 291–304.
- [19] A. Petrovic, L. Jovanovic, M. Antonijevic, N. Bacanin, M. Zivkovic, and J. Kaljevic, "Natural Language Processing of HTTP Content for Insider Threat Detection Optimized by Modified Metaheuristic," in *International Conference on Innovations in Cybersecurity and Data Science Proceedings of ICICDS*, Springer, 2024, pp. 299–314.
- [20] N. Bacanin et al., "Crop Yield Forecasting Based on Echo State Network Tuned by Crayfish Optimization Algorithm," *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Bangalore, India, 2024, pp. 1–6, doi: 10.1109/InC460750.2024.10649266.
- [21] S. Golubovic, A. Petrovic, A. Bozovic, M. Antonijevic, M. Zivkovic, and N. Bacanin, "Gold price forecast using variational mode decomposition–aided long short–term model tuned by modified whale optimization algorithm," in *International Conference on Data Intelligence and Cognitive Informatics*, Springer, 2023, pp. 69–83.



- [22] A. Petrovic, L. Jovanovic, N. Bacanin, M. Antonijevic, N. Savanovic, M. Zivkovic, M. Milovanovic, and V. Gajic, "Exploring metaheuristic optimized machine learning for software defect detection on natural language and classical datasets," *Mathematics*, vol. }, no. 18, p. 2918, 2024.
- [23] A. Petrovic, J. Perisic, L. Jovanovic, M. Zivkovic, M. Antonijevic and N. Bacanin, "Natural Language Processing Approach for Fake News Detection Using Metaheuristics Optimized Extreme Gradient Boosting," *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, Gwalior, India, 2024, pp. 252–257, doi: 10.1109/AIC61668.2024.10731062.
- [24] A. Vesic, M. Marjanovic, A. Petrovic, I. Strumberger, E. Tuba and T. Bezdan, "Optimizing Extreme Learning Machine by Animal Migration Optimization," *2022 IEEE Zooming Innovation in Consumer Technologies Conference (ZINC)*, Novi Sad, Serbia, 2022, pp. 261–266, doi: 10.1109/ZINC55034.2022.9840711.
- [25] M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, J. L. Raynier, G. Clowez, P. Boileau, and C. Ruetsch–Chelli, "Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis," *Journal of medical Internet research*, vol. 26, p. e53164, 2024.



ARTIFICIAL INTELLIGENCE IN THE CREATIVE INDUSTRY: STRATEGIC IMPLEMENTATION OF MARKETING TOOLS

Nemanja Jeličić*,
[0009-0004-2302-204X]

Ivana Vasić,
[0000-0002-3309-3472]

Jelena Gajić,
[0000-0001-9107-5398]

Aleksandar Mihajlović
[0009-0001-0777-2308]

Singidunum University,
Belgrade, Serbia

Abstract:

The emergence and development of Artificial Intelligence in the modern world caused the transformation of all marketing activities in companies around the world during the last decade. Revolutionary changes are particularly visible in the marketing activities of the creative industries, which according to available current data is one of the fastest growing sectors in the world. Artificial intelligence, as one of the most disruptive modern technologies, affects the development and implementation of marketing in creative industries, enabling advanced personalization and generation of new content, detailed market analysis and prediction of future trends, optimization of business processes, and the creation of autonomous creative solutions. Based on all of the above, the aim of the work is to explain how artificial intelligence is used in the marketing of creative industries, with special reference to the analysis of the tools used to create visual and textual content in advertising campaigns. The paper will also present and analyze case studies of successful marketing campaigns in which AI was applied. A special contribution of the paper is to provide insight into the future development and application of artificial intelligence in this area, with an emphasis on increasing innovation.

Keywords:

Artificial Intelligence, Creative Industries, Digital Marketing, Technology, Digital Tools.

INTRODUCTION

Artificial intelligence (AI) encompasses a wide range of data processing methods and techniques that enable machines and computers to imitate human intelligent behavior [1]. This field is interdisciplinary and includes various sub-disciplines such as computer vision, sensors, robotics and language processing [2]. AI is closely related to machine learning and data science, as it allows systems to learn and improve their capabilities through data analysis. Instead of relying on human intuition, artificial intelligence enables deep data analysis and content generation, including texts, visuals, photos and multimedia materials, that are specifically tailored to specific target groups [3]. In recent years, many companies have been adopting numerous technological innovations and solutions in order to improve the user experience of their consumers and thereby achieve an increase in return on investment (ROI).

Correspondence:

Nemanja Jeličić

e-mail:

njelicic@singidunum.ac.rs





By successfully implementing Artificial Intelligence as well as tools based on this concept, marketers can gain far better insight and achieve a comprehensive understanding of their target group.

The results of a survey carried out by Weber Shandwick in 2016, published by KRC Research, show that 55% of marketing managers from China, Great Britain and the USA believe that Artificial Intelligence will have a more significant impact on marketing as well as communication and customer relations than social media. Also, available data indicate that marketing is currently the fourth largest industry in the use of Artificial Intelligence and the sixth largest producer of technology [4].

2. THE USE OF AI IN THE CREATIVE INDUSTRY FROM A MARKETING ASPECT

Machine learning technologies and data analysis have become key to personalizing advertising and improving marketing campaigns, enabling brands to more effectively identify and reach their target audience. By using sophisticated algorithms, it is possible to create content that is aligned with the needs of end users, resulting in greater engagement and better business outcomes [3]. In the context of creative industries, these technologies enable the creation of innovative and dynamic marketing strategies, which include responsive visuals and interactive experiences, thus increasing the connection with the audience. Also, real-time data analysis enables quick adjustment of campaigns and optimization of content for the most relevant impact. Digital platforms, social media, and e-commerce are reshaping how cultural and creative products and services are marketed and presented to consumers. In this evolving landscape, digital marketing has become an essential tool that helps businesses maintain their competitiveness and thrive in an ever-more complex digital market environment. [5].

In the further part of the work, key aspects of the application of artificial intelligence (AI) will be explored in the analysis of consumer data and the prediction of their behavior, as well as in the generation of visual and textual content.

2.1. AI IN CONSUMER DATA ANALYSIS AND USER BEHAVIOR PREDICTION

One of the most important aspects of the application of AI in the field of marketing is its ability to, with the application of numerous tools and algorithms, collect and analyze a large amount of information about con-

sumers, including their demographic, economic, psychographic and other characteristics, but also their emotional reactions based on interaction with brands, which further contributes to the improvement of content personalization. Constant improvement and modification of existing algorithms results in each consumer becoming a unique segment over time. This progress brings numerous benefits to consumers - they receive messages and offers tailored to their interests and companies that can communicate with them in a much faster and more efficient way, which increases their engagement and loyalty [6]. The importance of content personalization is particularly highlighted in the Creative Industry sector, where by using numerous algorithms for data analysis, the mentioned sector can be very flexible and more precisely proactive. Flexibility is reflected in the ability to quickly and easily adapt to changes in consumer behavior, which enables the development of products and services that match their current needs. It also enables the creation of marketing campaigns that better connect with current social and cultural trends.

2.2. USING AI TO CREATE AUTONOMOUS CREATIVE SOLUTIONS

Creating unique original creative solutions is important in many industries including marketing, art, design, etc., where AI has a very important role. Unlike traditional creative processes where human inspiration and intuition were of particular importance for creating something new, with the help of modern technology, AI independently generates and manages creative products such as video materials, images, music, etc. This approach allows creative people, especially people from the world of marketing, to leave the so-called comfort zone and to approach creative and unconventional methods in their work.

Artificial intelligence is playing an increasingly significant role in enhancing user experience, with some of the most prominent technological solutions including chatbots, personalized and predictive content, automated content generation, and image recognition systems. These innovations enable more precise customization to user needs, improve service quality, and facilitate interaction between companies and users [7].

A wide range of tools are available in various segments of marketing and we can divide them into several types such as tools for: content generation and visual marketing, advertising, data analysis, data personalization, customer support, etc.



2.3. CONTENT GENERATION

In the world of text content creation, tools like ChatGPT, Jasper AI, and Adzoooma stand out, using natural language processing (NLP) models to create content for social media, ads, campaigns, and more. The essence of the work of these tools is based on entering short guidelines, also known as "prompts", which the tool interprets and uses as a basis for creating content [8].

In the creative industry, this way of working creates space for quickly creating content for social networks, developing slogans, writing synopses for advertisements, and even creating narratives for fashion campaigns or promotional films.

ChatGPT is based on the GPT-4 model developed by the OpenAI company, it stands out for its flexibility and ability to conduct an interactive dialogue. The user can request the tool to suggest alternative versions of the content, additional explanations or to change the writing style. This means creatives can play with variations, change the tone of communication or test multiple options before making a final decision. In creative teams, it is often used for brainstorming, creating texts for scenarios, campaigns, narration in video content or even designing communication with users through chatbots.

On the other hand, Jasper AI is specially adapted for the field of marketing and content that has a clearly defined business goal. It offers pre-designed templates covering Google ads, Facebook ads, product descriptions, blog posts and similar formats. In practice, it is most often used to design promotional messages, create SEO-friendly portfolio descriptions, run campaigns for new collections or create text frames for web pages by creative agencies, artists or designers. The user has the ability to determine the tone of communication - from formal to persuasive - and to generate text based on this information that can be directly used or further adjusted.

2.4. VISUAL MARKETING

Visual marketing is undergoing a significant transformation thanks to the development of modern digital tools such as DALL-E, Adobe Firefly and Canva Ai. These tools allow users to get ready-made illustrations, photos, visuals and animated content based on simple text descriptions. The creative industry, which until yesterday depended on the requisite skill and expertise in tools like Photoshop and Illustrator, now gets the opportunity to materialize ideas in seconds.

DALL-E, which was also developed by OpenAI, has the ability to generate photos for users based on prompts. This functionality is especially prevalent in the creative sector - from artists and designers to publishers and marketing agencies - who increasingly use these tools to create project sketches, posters or visual elements for social networks.

Adobe Firefly is integrated with Adobe tools such as Photoshop and Illustrator. It offers features such as generative fill and image creation based on prompts (Generative Fill and Text-to-Image). This allows creators not only to quickly generate images, but also to finish them professionally. The use of this functionality comes with a subscription fee, but it significantly speeds up the process for professional teams to more efficiently create presentations, mockups and working versions of projects.

Canva AI makes design accessible to everyone. With the help of functions such as Magic Design and Magic Write, the user enters a text description, and the platform automatically offers ready-made solutions. In the creative industries, this means that artists, influencers and small brands can independently create visuals based on text descriptions, logos, social media posts, portfolio presentations, without the need to hire professionals.

2.5. DATA ANALYTICS

In the contemporary digital landscape, where personalized communication sets the tone and every marketing move is measured and analysed, creativity without insight often remains untapped potential—making data-driven insights, supported by AI technologies that automate analysis, predict user behaviour, and optimize content in real time, essential to the success of modern campaigns [9].

Google Analytics 4 (GA4), which is one of the most prevalent tools in this domain, uses AI-based algorithms providing a more detailed insight into user behavior in order to recognize behavioral patterns, forecast conversion probability and analyze drop-off points. For teams in the creative industries, this means campaigns don't rely on intuition, but on concrete insights into what content holds attention and how to adjust the tone or format of content for better results [10].

HubSpot, as a CRM platform, offers marketing analytics with the help of artificial intelligence. In the creative industry, it is used to track customer behavior, audience segmentation and get suggestions for content optimization. Its simple interface is ideal for small teams, while more advanced options require paid packages.



Pecan AI is a tool that stands out because it allows teams to create predictive models without programming knowledge. Thanks to this, teams can create models such as sales forecasts, user behavior or campaign effectiveness. Pecan AI provides support not only in the analysis of results, but also in strategic planning.

2.6. SOCIAL NETWORKS

Managing social networks requires a constant rhythm, adapting to trends and understanding algorithms. Artificial intelligence makes this job much easier — it analyzes engagement, suggests posting times, and even automatically creates content.

As one of the most popular social media management tools, Hootsuite uses AI to identify the most effective moment to post, recognize posts that generate the most engagement, and schedule content based on audience behavior. In the creative industry, it is used to manage profiles of artists, brands and events across multiple platforms, as well as to plan campaigns, schedule posts and track results.

2.7. CUSTOMER SUPPORT

In a digital environment where users expect precise and quick answers, the quality of the user experience is a key factor in building trust and loyalty. Today, these expectations cannot be met without relying on technologies that enable real-time communication, where AI occupies a special place.

Drift is a tool that represents the concept of conversational marketing that successfully combines automation and personalization. It does not function like a classic chatbot, but represents a system that recognizes the user's intentions, asks additional questions and provides the necessary information. This kind of tool is particularly important and is widely used in the creative industry sector, which relies on digital communication and personalized access to users.

3. APPLYING AI IN CAMPAIGNS: EFFECTIVE EXAMPLES AND RESULTS

In the modern digital age, where speed, precision and personalization are expected, AI becomes a key ally of marketing teams. The implementation of AI itself goes beyond content automation as today it has the ability to shape entire campaigns, influence user experience and help brands better connect with audiences. Many brands are already using advanced AI-based solutions.

One of the most prominent examples is the Coca-Cola company, which used AI to analyze large amounts of data from social networks. The goal was to find out which emotions and themes users most associate with their products. Based on the results, a personalized marketing campaign was developed with messages and visuals adapted to different target groups. In 2023, the company presented the "Create Real Magic" project, which allows users to create unique visuals and illustrations with the help of AI, additionally including them in the process of communication with the brand [11].

In the fashion industry, the H&M brand uses AI for inventory management and personalization of online offers. Algorithms analyze user behavior and, based on that, suggest products to users that match their style and interests. This not only improves the user experience, but also increases the likelihood of purchase. This approach also contributes to the optimization of logistics, because the supply is automatically adjusted to the demand in real time.

The Starbucks company has developed a system called Deep Brew, which with the help of AI enables personalization of recommendations within the mobile application, but also helps in business organization through optimization of the company's employee schedule, inventory management and demand forecasting in different locations. The result is more efficient internal work and a better user experience.

Similarly, Netflix uses AI not only for personalized content recommendations, but also for promotional campaigns. Based on user viewing habits, the company generates personalized visuals and trailers that are tailored to individual interests. This approach significantly affects user engagement and therefore increases the likelihood that the user will choose a certain content [12].

In the cosmetics industry, Sephora has developed a Virtual Artist app that uses AI to enhance the user experience. The app allows users to virtually try on different products using augmented reality (AR) technology, while AI analyzes their choices and habits and offers personalized recommendations based on that, providing users with an advanced and interactive shopping experience.



Spotify is one of the most popular music streaming platforms. The company uses AI to analyze user habits and create personalized playlists. AI analyzes data about user listening, habits and behavior to recommend new songs and artists, increasing user engagement and satisfaction [13].

On the domestic market as well, there is an increasingly pronounced interest in the application of AI in the field of digital marketing and creative industry. Although the market is still evolving in terms of the integration of advanced technologies, certain companies are already using advanced technologies to analyze data, personalize content and improve the user experience.

One prominent example is the company AI Serbia, known for innovations in digital communication. AI has integrated AI within its chatbot services for customer support (which uses natural language processing technology) and within digital marketing. The virtual assistant allows users to receive relevant information about services, package suggestions according to personal needs, as well as answers to the most frequently asked questions. This approach enables a high degree of automation as well as a sense of personalized communication, which affects the quality of the user experience.

Banca Intesa has implemented AI in the area of risk management, but also for the automation of key processes through the RPA (Robotic Process Automation) system [14]. The Company also uses AI to segment customers and personalize offers, relying on data on customer transactional behavior and product usage. This approach allows the bank to improve marketing efficiency and build deeper relationships with clients [15].

Nordeus, a Serbia-based company best known for its mobile game Top Eleven, has attracted a global user base numbering in the millions. According to publicly available sources, the game produces around one terabyte of data per day, with the total data archive currently exceeding two petabytes. This extensive dataset captures a wide range of user interactions, from behavioral patterns and responses to game features, to levels of engagement with personalized content. The company has established a specialized team focused on data engineering and analytics. This team applies advanced analytical methods, including artificial intelligence algorithms, not only to process and interpret data but to translate insights into concrete actions. The use of AI allows for the identification of user preferences and behavioral trends, which in turn inform key decisions related to offer personalization within the game, user segmentation, and the refinement of marketing strategies [16].

4. CONCLUSION

Modern marketing and creative industries today increasingly rely on advanced digital technologies, which implies the use of AI in business. The application of AI includes a wide range of activities such as analyzing large amounts of data, predicting user behavior, automating processes, creating personalized content and improving user experience.

Global companies have been including AI in their business strategies for years, in order to improve the relationship with users and clearly position themselves in the market. Their examples show that AI does not diminish the importance of creativity, but adds additional values.

Similar developments are also observed in the domestic market, where more and more companies recognize the potential when it comes to the use of AI as a tool for business optimization and shaping communication that is aligned with the needs of users.

In the context of the creative industry, tools using generative models – such as DALL-E, Adobe Firefly or Canva AI – have transformed the way visuals, campaigns and concepts are developed. At the same time, analytics and automation tools, such as Google Analytics 4, HubSpot or Pecan AI, allow creative decisions to be based on real data and user behavior, rather than solely on intuition.

At a time when reaction speed, message relevance and user experience are of crucial importance, AI represents a key advantage over traditional work methods. Its role is no longer auxiliary - it is strategic. Organizations that recognize this in time will be not only more competitive, but also more prepared for the challenges of a dynamically changing market.



REFERENCES

- [1] A. Ng, "AI for Everyone," 2024. [Online]. Available: <https://www.deeplearning.ai/>.
- [2] V. Kotu and B. Deshpande, Data Science Concepts and Practice, Cambridge: United States: Elsevier Inc, 2019.
- [3] J. Gajić and A. Mihajlović, Advertajzing, Belgrade: Singidunum University, 2025.
- [4] J. Jöhnk, W. Malte and K. M. Wyrski, "Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors," *Business & Information Systems*, pp. 5-20, 2021.
- [5] I. Vasić and J. Gajić, "The importance of social media in the promotion and development of the creative industries in the Republic of Serbia," *The European Journal of applied economics*, pp. 1-18, 2018.
- [6] L. Ma and B. Sun, "Machine learning and AI in marketing – Connecting computing," *International Journal of Research in Marketing*, pp. 481-504, 2020.
- [7] P. Van Esch and J. Stewart Black, "Artificial intelligence (AI): revolutionizing," *Australasian Marketing Journal*, pp. 199-203, 2021.
- [8] I. Kofler and M. El Moussaoui, "AI's influence on the Creative and Cultural Industries," *Im@go*, pp. 291-312, 2024.
- [9] A. Mihajlović, J. Gajić and T. Papić, "Analysis of the Cost-Effectiveness of the University Instagram Marketing Campaign Using A/B Testing," *Sinteza*, p. 266–273, 2024.
- [10] R. Mohaimin, "The Influence of Google Analytics on E-commerce: Enhancing Customer Insights and Business Performance," *Journal of Business and Management Studies*, pp. 91-103, 2025.
- [11] J. Fong, "'Create Real Magic' Contest: Coca-Cola's AI-Powered Journey into Creativity," 29 November 2023. [Online]. Available: <https://medium.com/@janicefong/create-real-magic-contest-coca-cola-s-ai-powered-journey-into-creativity-ab6149e2cfcc>.
- [12] H. Steck, L. Baltrunas, E. Ehtsham, D. Liang, R. Yves and J. Basilico, "Deep Learning for Recommender Systems: A Netflix Case Study," *AI Magazine*, 2021.
- [13] C. Maheshwari, "Music Recommendation on Spotify using Deep Learning," *arXiv*, 2023.
- [14] B. Urošević, "Od sprečavanja zločina do procene rizika - kako srpske banke koriste AI," 18 September 2023. [Online]. Available: <https://rs.bloombergadria.com/tehnologija/inovacije/38333/srpske-banke-se-vec-oslanjaju-na-ai-a-sledi-jos-sira-primena>.
- [15] "How automation and digitization increased productivity by 25% at Banca Intesa Serbia," [Online]. Available: <https://bfconsulting.com/masteraps-at-banca-intesa/>.
- [16] PCPress, "Nordeus: Igre sa podacima," 5 November 2023. [Online]. Available: <https://pcpress.rs/nordeus-igre-sa-podacima>.



THE ECONOMICS OF AI-POWERED CALL CENTER DEVELOPMENT USING CHATGPT FOR THE NEEDS OF AN AUTOMOTIVE RETAIL BUSINESS

Pavle Dakić^{1,2*},
[0000-0003-3538-6284]

Tjaša Heričko²,
[0000-0002-0410-7724]

Živana Kljajić⁴,
[0000-0001-7510-6852]

Vladimir Todorović⁵
[0000-0001-9794-9527]

¹Singidunum University,
Belgrade, Serbia

²Faculty of Informatics and
Information Technologies,
Slovak University of Technology
in Bratislava,
Bratislava, Slovakia

³Faculty of Electrical Engineering and
Computer Science,
University of Maribor,
Maribor, Slovenia

⁴Faculty of Business Economics,
Pan-european University Apeiron,
Banja Luka Bosnia & Herzegovina

⁵Faculty of Business Studies and Law,
MB University,
Belgrade, Serbia

Correspondence:

Pavle Dakić

e-mail:

pdakic@singidunum.ac.rs

Abstract:

Amid the shift toward automation in customer service through artificial intelligence (AI), this research study explores the practical integration of OpenAI Realtime API using the GPT-4o-mini-realtime-preview-2024-12-17 model into an AI-powered call center tailored for the local automotive webshop. Based on OpenAI's documentation, this study investigates the cost structure, technical implementation, and integration techniques of a ChatGPT language model through ChatGPT Realtime API, a streaming API suitable for real-time interactions with token-by-token response and low latency within a PHP-based environment. The study aimed to develop a scalable, multi-lingual AI-powered call center prototype leveraging WebRTC and PHP to deliver online, i.e., text-based, and phone, i.e., speech-based, customer assistance. The system employs WebRTC for real-time speech and text communication, while PHP facilitates seamless interaction with existing databases and backend systems. The prototype focuses on system integration, real-time data processing, and efficient API connectivity to improve automation in the context of online auto parts retail. The resulting prototype represents a cost-effective alternative to traditional call centers, offering faster response times, reduced operational expenses in terms of staffing expenses, and improved customer experience.

Keywords:

AI Call Center, OpenAI Realtime API Integration with PHP, GPT-4o-mini-realtime-preview-2024-12-17, Customer Service Automation, AI Development Costs.

INTRODUCTION

The development of call centers using artificial intelligence (AI) has been causing great interest at the global level in the past few years. Of particular focus is the transformative impact of AI on customer service and the call center industry, as confirmed by recent Stanford University research on generative AI technologies [1]. The study's findings indicate measurable improvements in business efficiency driven by AI integration. However, the aforementioned trend has also raised concerns, particularly regarding potential job losses in developing countries that rely heavily on call center outsourcing. As AI-supported systems increasingly cover customer support functions, businesses with call centers, especially business process outsourcing (also known as BPO), can be expected to undergo key economic changes, redefining the way of thinking and businesses inner workings [2].



To address these trends, ChatGPT-4o-realtime is among the most current state-of-the-art AI models that allow organizations to fully automate customer interactions with rapid responses, significantly reducing wait times and improving service quality [3]. Various industries, such as the automotive sector, can exploit its capabilities for text generation, contextual background information retention for product-specific queries, and real-time language translation in the case of international customers. Human interaction and meeting business needs correlate with software design patterns that influence the process of creating AI agents [4].

Despite its prospects, setting up AI-supported call centers is not a trivial task and presents numerous challenges. To name a few, high-performance computing infrastructure and comprehensive data engineering pipelines are necessary, which implies continuous fine-tuning to ensure and maintain the accuracy and ethical behavior of AI applications across several different industrial branches [5, 6]. Current obstacles include substantial development costs and integration difficulties with existing edge computing, the Internet of Things (IoT), and broader IT ecosystems [7, 8]. Moreover, issues of data privacy, cybersecurity, and the risk of external exploitation or system abuse by third parties must be proactively addressed [9, 10]. Another limitation lies in the inability of AI systems to replicate the emotional intelligence intrinsic to human agents, which may affect user satisfaction in certain contexts.

The increasing automation of service functions introduces the possibility of substituting human labor, which leads companies to strive to establish a balance between financial costs and new technologies, but also doubts about the reform of the labor force and the code of ethics [11].

This paper investigates the economic, technological, and societal implications of real-time AI models, specifically GPT-4o and the gpt-4o-mini-realtime-preview-2024-12-17 model, within the context of AI-driven call centers. The research study focuses on the performance and impact of API capabilities on future user engagement trajectories. The paper covers the feasibility of creating a prototype capable of real-time communication and direct interaction with customers using predefined instructions. The use case of the proposed system is in the automotive industry sector, more specifically serving as a dynamic call agent, handling inquiries and service requests for online auto parts purchases.

The present study provides valuable insights into the feasibility, technical implementation, and economic impact of incorporating AI-powered automation into customer support operations. By designing and evaluating real-time, GPT-4-powered, API-based AI call centers, this research highlights key advancements in AI-driven interactions with customers. The following contributions emphasize the importance of this research:

1. *Understanding of existing solutions written in JavaScript.* A review of current implementations reveals that most development of AI call center prototypes relies heavily on JavaScript technologies, particularly Node.js and the React framework. While effective for rapid prototyping, these technologies are often not best suited for large-scale enterprise adoption due to server architectural complexity, limited integration with legacy systems, and costs related to that [12].
2. *Development of PHP-based solution for real-time communication with ChatGPT via WebRTC.* This study presents a scalable and customizable platform that facilitates the integration of AI-driven automation within existing enterprise IT infrastructures. This contribution is presented in Section 4.1, where we describe how PHP and ChatGPT API connectivity with WebRTC were combined for real-time audio and text communication. The above-mentioned bridges AI capabilities with web-based customer service solutions, offering enterprises an efficient way to enhance their support systems.
3. *Assessment of economic viability and costs.* This research evaluates the financial implications of utilizing GPT-4's real-time API, including API usage fees, infrastructure costs, and potential operational savings. The findings, detailed in Section 4.2, outline the expenses associated with implementing a ChatGPT-powered call center solution.

Based on the stated objectives, this research study aims to develop a GPT-4-powered real-time AI call center utilizing PHP and WebRTC technologies, with a specific focus on the integration of existing enterprise IT infrastructures for automated customer support. The research adopts a structured approach, beginning with an overview of the selected technology stack, followed by system design, cost evaluation, and practical implementation considerations.



The organizational structure of the paper is as follows: Section 1 covers the basic characteristics of the challenges we address and face during research, including the core research problem and the obstacles in constructing AI-powered call centers. Section 2 presents a brief review of the literature, addressing call centers and their operational sustainability, as well as the regulatory frameworks governing their use. Section 3 presents the employed research method and the scope of the investigations. Section 4 presents the achieved scientific research results and contributions in the aspect of creating a PHP-based implementation of an AI call center and evaluating its economic feasibility during its use. Section 5 offers a discussion and summarizes the findings concerning the research questions. Finally, Section 6 concludes the paper by summarizing the viability of integrating GPT-4, PHP, and WebRTC technologies for developing AI-powered call centers, showcasing the potential for increased automation, cost savings, and improved customer service efficiency.

2. LITERATURE REVIEW

The literature review section briefly highlights some of the difficulties encountered by other researchers while investigating AI and its application to call centers. The impact of AI in education on fundamental human rights is examined in an EU scoping study by Berendt et al. [13]. As highlighted by Fukunaga et al. [14], who developed a staff system for scheduling called Director for customer service centers where the economics of AI call centers cover the cost-effectiveness of starting up in the EU. The topic has gained close attention in the past few years as one of the interesting options in solving the problem of multiple language support. While the second challenge in staff scheduling for call centers is critical, it also allows the other parts of the organization to be more efficiently structured, which may have previously been overlooked.

The financial viability of AI call centers in the EU has received a lot of interest in recent years. As highlighted by Fukunaga et al. [14] in the study, they emphasized the importance of staff scheduling in call centers. For this, they suggested the creation of a director and a staff scheduling system for contact centers. In the case of outsourcing call centers, as pointed out by Morgan et al. [15], other nations like India have raised concerns linked to the globalization of economics and linguistic policies. While Berendt et al. [13] conducted a study and reflected on the influence of AI educational institutions on fundamental human rights.

2.1. CONTEXT OF THE EUROPEAN UNION AND AI CALL CENTERS

According to Vachudova et al. [16], the European Union's accession process has impacted party competitiveness and alignment in economic and cultural areas. Samaniego-Ocaña et al. [17] describe the evolving models for managing emergency calls in EU dispatch coordination centers, noting changes in law and response times.

2.2. SUSTAINABLE AI REGULATION AND EU AI ACT

Sustainable AI regulation, which is covered by Philipp Hacker [18], advocates transparency mechanisms and sustainability-by-design principles. Kraus et al. [19] stress the need for a multidisciplinary approach when integrating AI technology in customer service, combining human operators with virtual agents. On the other hand, Gao et al. [20] address the development of AI approaches, such as multiturn dialogue agents, to aid human salespeople in contact centers. Overall, the literature indicates an increasing interest and investment in AI technology in EU call centers.

3. MATERIALS AND METHODS

The research framework employed in this work is based on two targeted research questions that guided the software solutions design process, ensuring that our investigations stayed focused on our objectives. These questions were designed to provide in-depth information about the feasibility, economic viability, and technological challenges of developing a GPT-4-powered, real-time, API-based AI call center prototype that connects to a proprietary database and serves the auto industry and online parts retailers. Based on this research setting, the following study areas were identified:

1. RQ1. How to create a ChatGPT AI-powered call center in a PHP environment?
2. RQ2. What are the economic benefits and cost implications of implementing ChatGPT models for call center operations?

The research questions formulated in this research directed the study investigations and guided the development of a prototype solution to address identified gaps in AI-driven customer service. Hence, the primary objective of the proposed prototype is to leverage state-of-the-art AI functionalities to streamline customer interactions, analyze specific questions, and efficiently extract relevant information from its specific database.



We hypothesize that offering context-aware responses might improve customer service and increase overall service quality. The innovative system could meet the specific needs of the automotive and e-commerce industries, setting completely new standards for automated customer service in a competitive market.

4. RESULTS

The research results indicate that monthly personnel costs in the call center can be significantly reduced, allowing businesses to shift their focus toward distribution and delivery operations. The prototype demonstrates how a ChatGPT real-time API-based call center can effectively automate customer support, specifically for an online auto parts import and sales business,

improving efficiency, reducing costs, and enhancing the user experience.

4.1. APPLICATION OF THE REAL-TIME API USING PHP

Code Listing 1 illustrates key parts of the implementation responsible for routing the obtained information. This component plays a central role in directing processed data to the appropriate communication channels within the AI-driven call center system.

Figure 1 presents three variations of the interactive user interface, through which users can communicate through text, voice, and video interactions. Within this environment, the AI assistant is capable of processing user-provided images to deliver detailed information on auto parts.

```
<?php
require 'vendor/autoload.php';
// code part
$DEFAULT_INSTRUCTIONS = "Introduce yourself to the user by saying your name.
Prof. dr Zoran Radmilovic, you are a helpful Service Advisor.
You are fluent in Serbian, English, Slovak, and Slovenian.
You have extensive experience in the automotive business.
During the phone contact, introduce yourself and greet the client.
You may find the newest service pricing by searching the website
https://ovlasceniservis.com/.
When the user inquiries about the company's operating hours,
you can provide the following information: Authorized Service doo ";
$app->post('/rtc', function (Request $request, Response $response) use ($DEFAULT_
INSTRUCTIONS) {
    $OPENAI_API_KEY = $_ENV['OPENAI_API_KEY'];
    // Get the raw SDP body
    $body = $request->getBody()->getContents();
    // Construct the OpenAI URL with parameters
    $url = 'https://api.openai.com/v1/realtime';
    $url .= '?model=' . urlencode('gpt-4o-mini-realtime-preview-2024-12-17');
    $url .= '&instructions=' . urlencode($DEFAULT_INSTRUCTIONS);
    $url .= '&voice=echo'; // Voice model style
    // Initialize cURL session
    $ch = curl_init($url);
    // code part
```

Listing 1. Code Snippet for to initialize routing DEFAULT INSTRUCTIONS in PHP. *Source:* author's contribution

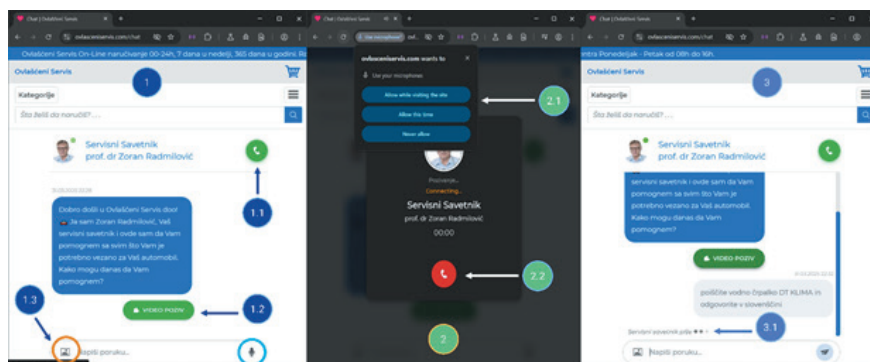


Figure 1. Interactive chat environment with support for text, audio and video communication. *Source:* author's contribution



5. DISCUSSION

The preliminary research results demonstrated that a GPT-4-driven AI call center prototype based on real-time APIs has considerable benefits in terms of operational efficiency, linguistic support, and scalability. Key findings indicate that the experimental model has adequate potential for future development and can serve as the foundation for the development of business solutions. Despite its benefits, the implementation of such a system still faces several notable challenges. One key challenge lies in the high initial costs and considerably challenging integration with existing active information systems. This is especially true for businesses in competitive industries, such as online automotive retail, where operational systems tend to be rigid and costly to change and upgrade. Based on the previously posed research questions, key findings may be summarized as follows:

- RQ1. We successfully implemented a prototype AI-powered call center using ChatGPT in a PHP-based environment by integrating Webrtc for real-time voice and text communication, along with the GPT-4o-mini-realtime-preview-2024-12-17 model. The implemented prototype demonstrated the GPT-4 API in action for AI low-latency responses and a PHP backend for smooth communication, consequently increasing efficiency with session management, queuing, and tracking user requests in automotive retail customer support.
- RQ2. Our findings suggest that using ChatGPT in its experimental mode as the core building block for an AI-powered call center can significantly reduce personnel costs, improve response times, and boost system scalability. Despite hefty early integration costs and an experimental model of the language model, the implemented solution demonstrated the advantages of automation with AI-powered solutions that can be more cost-effective than traditional call centers from a long-term perspective.

6. CONCLUSION

This paper emphasizes the promising potential of a GPT-4-powered API-based AI call center (using the gpt-4o-mini-realtime-preview-2024-12-17 model) for real-time applications in the automotive industry and online auto parts retail while pointing out important constraints, such as development costs and complex system integration.

PHP, with its extensive usage as a web technology, represents a powerful environment for building scalable solutions that support real-time communication while remaining compatible with legacy systems. In our case, the integration within a PHP-based environment was essential to link advanced AI capabilities to existing web infrastructure, enabling effective API connectivity and dynamic data handling. However, it should be emphasized that if not implemented carefully, such systems can negatively impact customer satisfaction, potentially leading to abandoned purchases or reduced customer engagement.

Future research should prioritize the iterative development of AI-based call centers with extensive real-world testing with end-users, i.e., customers, with an additional focus on refining PHP implementations and integrations with existing systems and further optimization of the use of AI to maximize benefits while minimizing operational obstacles.

7. ACKNOWLEDGEMENTS

The work reported here was supported by Erasmus+ ICM 2023 No. 2023-1-SK01-KA171-HED-000148295 and Model-based explication support for personalized education (Podpora personalizovaného vzdelávania explikovaná modelom) - KEGA (014STU-4/2024).

REFERENCES

- [1] S. University, "Will Generative AI Make You More Productive at Work? Yes, But Only If You're Not Already Great at Your Job. | Stanford HAI," Stanford, April 2023. [Online]. Available: <https://hai.stanford.edu/news/will-generative-ai-make-you-more-productive-work-yes-only-if-youre-not-already-great-your-job>.
- [2] E. Brynjolfsson, D. Li and L. Raymond, "Generative AI at work," *The Quarterly Journal of Economics*, 2025, DOI: <https://doi.org/10.1093/qje/qjae044>.
- [3] F. Cák and P. Dakić, "Creating Feature Model for YAML Generator in CI/CD Pipelines with React Web Application," in *Proceedings of Ninth International Congress on Information and Communication Technology*, Springer Nature Singapore, 2024, DOI: https://doi.org/10.1007/978-981-97-3305-7_42, p. 529–539.
- [4] M. Daud Haiderzai, I. Ul Haq Safi and P. Dakić, "Exploring the Impact of Technology on Human Interaction and Engaging Business Needs Through Software Design Patterns," *JITA - Journal of Infor-*



- tion Technology and Applications (Banja Luka) - APEIRON, vol. 24, October 2024, DOI: <https://doi.org/10.7251/jit2402117h>.
- [5] P. Dakić, "Importance of knowledge management for CI/CD and Security in Autonomous Vehicles Systems," *JITA - Journal of Information Technology and Applications (Banja Luka) - APEIRON*, vol. 24, May 2024, DOI: <https://doi.org/10.7251/jit2401007d>.
 - [6] OpenAI, "Realtime API - OpenAI API," OpenAI, March 2025. [Online]. Available: <https://platform.openai.com>.
 - [7] P. Dakić, "Software compliance in various industries using CI/CD, dynamic microservices, and containers," *Open Computer Science*, vol. 14, January 2024, DOI: <https://doi.org/10.1515/comp-2024-0013>.
 - [8] T. E. Ali, F. I. Ali, P. Dakić and A. D. Zoltan, "Trends, prospects, challenges, and security in the healthcare internet of things," *Computing*, vol. 107, December 2024, DOI: <https://doi.org/10.1007/s00607-024-01352-4>.
 - [9] P. Dakic, M. Zivkovic, L. Jovanovic, N. Bacanin, M. Antonijevic, J. Kaljevic and V. Simic, "Intrusion detection using metaheuristic optimization within IoT/IloT systems and software of autonomous vehicles," *Scientific Reports*, vol. 14, October 2024, DOI: <https://doi.org/10.1038/s41598-024-73932-5>.
 - [10] P. Dakić, I. Stupavský and V. Todorović, "The Effects of Global Market Changes on Automotive Manufacturing and Embedded Software," *Sustainability*, vol. 16, p. 4926, June 2024, DOI: <https://doi.org/10.3390/su16124926>.
 - [11] M. D. Haiderzai, P. Dakić, I. Stupavský, M. Aleksić and V. Todorović, "Pattern Shared Vision Refinement for Enhancing Collaboration and Decision-Making in Government Software Projects," *Electronics*, vol. 14, p. 334, January 2025, DOI: <https://doi.org/10.3390/electronics14020334>.
 - [12] T. Heričko, "Analiza in primerjava uporabe PHP in Node.js pri razvoju spletnih strani," 2017. [Online]. Available: <https://dk.um.si/IzpisGradiva.php?lang=slv&id=67222>.
 - [13] B. Berendt, A. Littlejohn and M. Blakemore, "AI in education: learner choice and fundamental rights," *Learning, Media and Technology*, vol. 45, p. 312–324, July 2020, DOI: <https://doi.org/10.1080/17439884.2020.1786399>.
 - [14] A. Fukunaga, E. Hamilton, J. Fama, D. Andre, O. Matan and I. Nourbakhsh, "Staff Scheduling for Inbound Call and Customer Contact Centers," *AI Magazine*, vol. 23, p. 30, March 2002, DOI: <https://doi.org/10.1609/aimag.v23i4.1667>.
 - [15] B. Morgan and V. Ramanathan, "Outsourcing, Globalizing Economics, and Shifting Language Policies: Issues in Managing Indian Call Centres," *Language policy*, vol. 8, p. 69–80, 2009, DOI: <https://doi.org/10.1007/s10993-008-9111-x>.
 - [16] M. A. Vachudova and L. Hooghe, "Postcommunist Politics in A Magnetic Field: How Transition and EU Accession Structure Party Competition on European Integration," *Comparative european politics*, p. 179–212, 2009, DOI: <https://doi.org/10.1057/cep.2008.9>.
 - [17] J. M. Samaniego-Ocaña and M. Granados-Martín, "Evolving Models for Handling Emergency Calls in Spanish and European Dispatch Coordination Centers," August 2021. [Online]. Available: https://revistaemergencias.org/wp-content/uploads/2023/08/Emergencias-2021_33_3_211-217.pdf.
 - [18] P. Hacker, "Sustainable AI Regulation," *Common Market Law Review*, vol. 61, 2024, DOI: <https://doi.org/10.54648/cola2024025>.
 - [19] S. Kraus, Y. Oshrat, Y. Aumann, T. Hollander, O. Maksimov, A. Ostroumov and N. Shechtman, "Customer Service Combining Human Operators and Virtual Agents: A Call for Multidisciplinary AI Research," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, 2023, DOI: <https://doi.org/10.1609/aaai.v37i13.26795>.
 - [20] W. Gao, X. Gao and Y. Tang, "Multi-Turn Dialogue Agent As Sales' Assistant in Telemarketing," in *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, DOI: <https://doi.org/10.1109/IJCNN54540.2023.10192042>.



ASSESSMENT OF WASTEWATER TREATMENT, WATER QUALITY, AND CLIMATIC STRESSORS ON FRESHWATER RESOURCES IN THE EUROPEAN UNION COMPARED TO SERBIA: TRENDS AND IMPLICATIONS

Aleksandra Mitrović^{1,2},
[0000-0002-5183-6276]

Nada Ratković Kovačević^{1*},
[0000-0001-6398-4391]

Ivana Ilić²,
[0000-0001-7364-8236]

Dejan Ilić³,
[0000-0001-8966-9955]

Maja Anđelković²
[0000-0002-5507-9831]

¹The Academy of Applied Studies
Polytechnic,
Belgrade, Serbia

²University "Union - Nikola Tesla",
Faculty of Information Technology and
Engineering,
Belgrade, Serbia

³University "Union - Nikola Tesla",
Faculty of Business Studies and Law,
Belgrade, Serbia

Correspondence:

Nada Ratković Kovačević

e-mail:

nratkovickovacevic@politehnika.edu.rs

Abstract:

The sustainable management of water resources remains critical to environmental protection, public health, and ecological stability within the European Union (EU). This study evaluates recent EU progress regarding wastewater treatment infrastructure, analyses water quality indicators including organic pollutants, nitrates, and phosphates, and examines climate-induced pressures, particularly drought events, affecting freshwater availability. Our assessment is based on statistical analyses from Eurostat and the European Environment Agency (EEA). Results demonstrate improvements in wastewater treatment connectivity, with EU average connections to secondary treatment systems rising significantly from 72.6 % in 2006 to 80.9 % in 2021. However, despite advances, water quality challenges persist, notably eutrophication driven by stable nitrate levels in groundwater (averaging 20.5 mg/l in 2021) and increasing phosphate concentrations in rivers (up by 21.3 % since 2016). Concurrently, climatic impacts, especially drought events, have exacerbated water stress, significantly affecting about 29 % of the EU territory periodically. Comparative data from Serbia reveals similar challenges but underscores significant gaps in wastewater treatment coverage and water quality, highlighting the importance of strategic regional and EU-wide water governance measures. Effective policy implementation and investment in sustainable water infrastructure are essential to counteract environmental degradation and ensure long-term freshwater security.

Keywords:

Wastewater Treatment, Water Quality, Eutrophication, Climate Change, Water Scarcity.

INTRODUCTION

The sustainable management of freshwater resources is crucial for environmental conservation, safeguarding public health, and ensuring long-term economic sustainability within the European Union (EU). Freshwater systems provide essential ecological services, support biodiversity, and underpin human activities ranging from agriculture to industry [1], [2], [3], [4]. However, these resources face mounting pressures from both anthropogenic activities and changing climatic conditions. Sustainable Development Goal 6 (SDG 6) explicitly emphasizes the necessity to guarantee the availability and sustainable management of water and sanitation for all, aligning closely with existing EU environmental frameworks such as the Water Framework Directive, the Urban Waste Water Treatment Directive, and related environmental policies designed to protect water quality and availability [2].



In recent decades, the EU has significantly advanced its policy efforts aimed at improving water resource management [5]. Despite these efforts, persistent and emerging threats, notably water pollution from organic and nutrient contaminants, continue to challenge environmental and public health goals. Nutrient pollutants, including nitrates and phosphates, primarily originating from agricultural runoff and inadequately treated wastewater, have intensified eutrophication, leading to the degradation of aquatic ecosystems. These challenges are further exacerbated by the accelerating impacts of climate change, mainly the increasing frequency and severity of drought conditions, which strain freshwater availability and quality [3], [4], [6], [7].

Additionally, a comparative perspective, such as that involving non-EU countries like Serbia, reveals both shared environmental issues and distinct regional disparities in water infrastructure and management effectiveness. Serbia and other countries in the region demonstrate significant gaps in wastewater treatment coverage and persistent water quality challenges, which underscore the necessity for targeted, strategic policy responses, investment in infrastructure, and regional cooperation to address transboundary water management challenges [5], [7].

Given these contexts, this paper evaluates EU progress and shortcomings regarding wastewater treatment, investigates current water quality trends focusing on key pollutants such as organic matter, nitrates, and phosphates, and explores the increasing climatic pressures that exacerbate water scarcity issues. The assessment draws insights from comprehensive statistical analyses to inform strategic recommendations for policy enhancement, sustainable infrastructure investments, and cross-border cooperation to secure freshwater resources sustainably into the future.

2. METHODOLOGICAL FRAMEWORK FOR EVALUATING WATER MANAGEMENT AND QUALITY

The methodology utilized in this paper involves a detailed analysis of comprehensive statistical datasets primarily provided by Eurostat and the European Environment Agency (EEA), known for their reliability and extensive coverage of environmental and sustainability data within the EU context. Several critical environmental indicators were carefully selected to provide a robust evaluation of water management effectiveness across the EU. The indicators include wastewater treatment

connectivity, specifically the percentage of populations served by secondary or advanced wastewater treatment systems, reflecting infrastructure capability and policy implementation effectiveness.

The analysis of comprehensive statistical datasets provided by Eurostat and the European Environment Agency forms served as the foundation of this research. Critical environmental indicators have been selected to evaluate the effectiveness of water management across the EU. Wastewater treatment connectivity, measured by the percentage of populations served by secondary or advanced treatment systems, was a key metric for assessing infrastructure capabilities and policy implementation. Higher connectivity rates are often linked to improved public health outcomes and reduced environmental risks, demonstrating the importance of robust wastewater management systems [8], [9].

Water quality assessment incorporates biochemical oxygen demand (BOD) alongside nitrate and phosphate concentrations. BOD levels indicate organic pollutant presence and reflect the operational efficiency of wastewater treatment facilities. Nitrate and phosphate measurements provide insights into nutrient pollution, which significantly contributes to eutrophication. Agricultural runoff, industrial discharges, and urban wastewater management practices influence these parameters, highlighting the complexity of maintaining water quality standards. Understanding the interactions between these factors is essential for addressing pollution sources and improving environmental protection strategies [10], [11], [12], [13].

Historical datasets covering the period from 2000 to 2021 facilitate the identification of long-term trends and variations in water quality and availability. Evaluating the impact of major policy directives, such as the Urban Waste Water Treatment Directive, helps track infrastructure advancements and their effectiveness in mitigating pollution. Longitudinal trend analyses and comparative benchmarking across EU member states reveal notable regional disparities [14]. Differences in wastewater treatment implementation across countries underscore the varying degrees of success in policy enforcement and infrastructure development.

Comparing these findings with data from Serbia highlights additional challenges outside the EU framework. Lower connectivity rates to advanced wastewater treatment and elevated nutrient loads in Serbian water bodies illustrate gaps in infrastructure and policy implementation. Economic constraints, regulatory differences, and geographical factors influence these variations, emphasizing the necessity of regional cooperation in addressing water management challenges [2].



Climatic stressors, such as droughts and water scarcity, further affect freshwater resource availability. Temperature fluctuations and changing precipitation patterns contribute to regional disparities in water supply and demand, complicating efforts to maintain sustainable water management. Integrating statistical evaluations with engineering and environmental analyses allows for a deeper understanding of the interactions among infrastructure, pollution control, and climate-related challenges. The combined examination of these factors provides a structured basis for assessing the effectiveness of water resource management across different regions [8], [11].

3. EVALUATION OF WASTEWATER TREATMENT, WATER QUALITY TRENDS, AND CLIMATIC IMPACTS

Evaluation of wastewater treatment, water quality trends, and climatic impact is a multifaceted subject that requires an integrated analytical approach combining environmental engineering, hydrology, and climate science. Wastewater treatment efficacy is fundamental to public health and environmental sustainability, as it directly influences the levels of organic and inorganic contaminants discharged into water bodies. Advanced treatment processes, particularly secondary and tertiary treatments, are critical in reducing biochemical oxygen demand and controlling nutrient levels, specifically nitrates and phosphates, which if unmitigated, lead to eutrophication and deterioration of aquatic ecosystems [9], [12], [13]. Concurrently, long-term water quality trends, derived from robust statistical analyses of historical datasets from sources such as Eurostat and the European Environment Agency, provide insight into the temporal dynamics of these contaminants. These trends

reveal the effectiveness of existing wastewater treatment infrastructures and highlight the need for continuous upgrades and policy interventions.

Climatic impacts further complicate this scenario by altering hydrological cycles and exacerbating water scarcity and quality issues. Shifts in temperature, precipitation patterns, and the frequency of extreme weather events, attributable to climate change, affect the dilution, dispersion, and concentration of pollutants in aquatic systems. Consequently, climatic variability not only stresses current water treatment capacities but also influences the performance of these systems over time, necessitating adaptive management strategies. Integration of climatic data with water quality monitoring facilitates a comprehensive understanding of how these external stressors interact with anthropogenic pressures, ultimately informing sustainable water resource management. This integrated evaluation underscores the imperative for a coordinated approach in upgrading wastewater treatment facilities, reducing waste generation and implementing climate resilience measures to ensure both environmental protection and public health in an era of rapidly changing global conditions [15], [16], [17].

The analysis indicates significant EU-wide progress in wastewater treatment infrastructure, with connectivity to secondary treatment facilities improving from 72.6 % in 2006 to 80.9 % in 2021. This marked improvement highlights the effectiveness of EU policy initiatives, particularly the Urban Waste Water Treatment Directive, which has mandated enhanced treatment standards. Despite these infrastructural advancements, critical water quality issues persist, especially concerning nutrient pollution or waste generation, in the EU and neighbouring countries as well [15], [16].

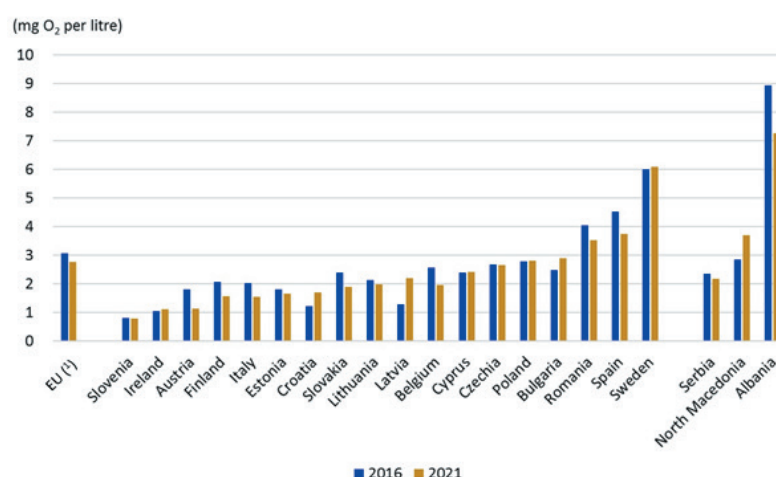


Figure 1. Biochemical oxygen demand in rivers, by country, 2016 and 2021 (mg O₂ per litre) [18]. (Fig. 6 in [18])

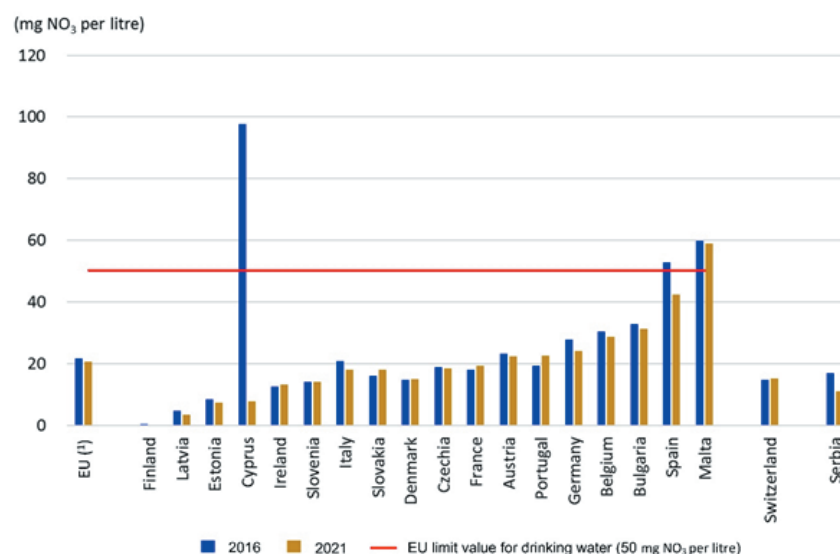


Figure 2. Nitrate in groundwater, by country, 2016 and 2021 (mg NO₃ per litre) [18]. (Fig. 8 in [18])

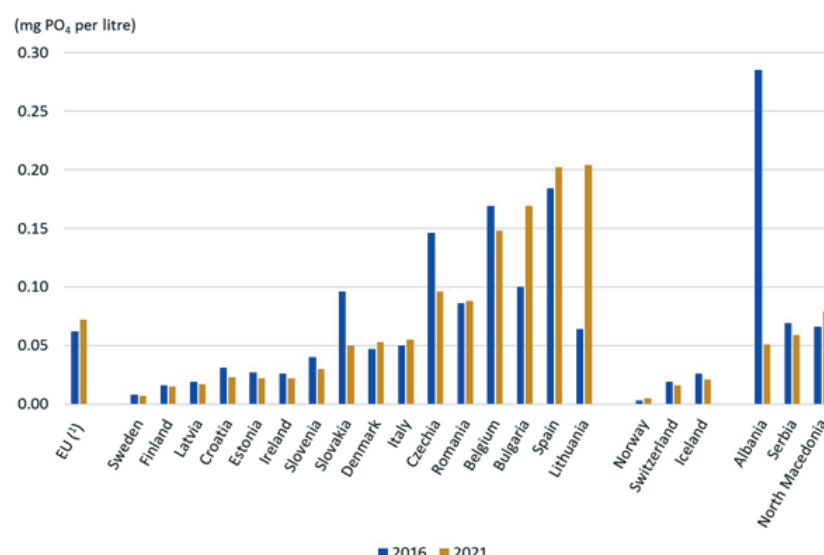


Figure 3. Phosphate in rivers, by country, 2016 and 2021 (mg PO₄ per litre) [18]. (Fig. 10 in [18])

The biochemical oxygen demand levels (O₂ [mg/l]), indicative of organic pollutants, have generally decreased (Figure 1 [18]), reflecting successful upgrades in wastewater treatment facilities and better industrial practices.

However, the persistent presence of nutrients—specifically nitrates and phosphates—remains a substantial environmental and public health concern [19]. Average nitrate concentrations in groundwater (NO₃ [mg/l]) have stabilized around 20.5 mg/l, continuously challenging the EU's goal to significantly reduce nutrient loads and prevent eutrophication (Figure 2 [18]).

Meanwhile, phosphate concentrations (PO₄ [mg/l]) in river systems have risen sharply, by 21.3 % since 2016, predominantly due to intensified agricultural activities, suboptimal wastewater treatment in rural and peri-urban areas, and inconsistent enforcement of environmental regulations (Figure 3 [18]).

The climatic impact on freshwater resources, notably through drought events, has exacerbated water scarcity conditions across the EU. Approximately 29 % of EU territory now experiences severe water stress periodically, with pronounced regional disparities influenced by climate variability, population density, agricultural



demand, and tourism. Countries in southern Europe, such as Cyprus and Malta, experience particularly high water exploitation indices, indicating unsustainable freshwater use. This finding underscores the critical need for adaptive management strategies and climate resilience planning, in EU and other countries, e.g. Ireland [15].

A comparative analysis involving Serbia highlights broader challenges that extend beyond EU borders, emphasizing pronounced gaps in wastewater treatment infrastructure and persistent nutrient pollution. Serbia's lower connectivity rates to advanced wastewater treatment and elevated nutrient loads in water bodies further stress the urgency of cross-border policy integration and regional cooperation.

Policy effectiveness across the EU varies, with notable successes tempered by ongoing implementation gaps. Enhanced enforcement of existing directives increased financial investments in infrastructure, particularly in rural and lower-density regions, and proactive climate adaptation measures are crucial for improving freshwater sustainability. Future policies should prioritize integrated water resource management, leveraging technological innovation and fostering cross-regional collaborations to address the intertwined challenges of water quality, climate change, and infrastructure deficits comprehensively [3], [4], [8], [20].

4. CONCLUSION

This study emphasizes the substantial progress achieved in the EU wastewater treatment infrastructure over recent decades, while simultaneously highlighting persistent challenges in maintaining water quality amid increasing climate-induced water stress. Significant enhancements in treatment connectivity and advanced processing methods have contributed to reductions in organic pollutants, yet issues such as eutrophication remain prevalent due to elevated concentrations of nutrients, specifically nitrates and phosphates. These findings indicate that while infrastructural developments have been successful in mitigating some aspects of water pollution, the management of nutrient loads continues to be a critical environmental concern.

Furthermore, the impacts of climate variability have intensified the stress on freshwater resources, as prolonged droughts and erratic precipitation patterns exacerbate water scarcity and quality degradation. The evidence suggests that adaptive management strategies must be integrated into current water resource poli-

cies to effectively address the dual challenges of nutrient pollution and climatic stress. Enhancing monitoring capabilities and predictive modelling is essential to better understand the temporal dynamics of pollutant dispersion and the influence of extreme weather events on water systems.

Recommendations emerging from this study advocate for increased investment in wastewater treatment infrastructure to expand capacity and improve operational efficiency. Moreover, the enforcement of environmental regulations requires strengthening to ensure consistent application of policies across different regions. Enhanced regional cooperation is also imperative, as coordinated efforts among EU member states and neighbouring regions can facilitate the sharing of best practices and the development of integrated water management strategies. Such collaborative approaches are vital for establishing resilient water systems capable of sustaining both ecological integrity and public health in the face of evolving environmental pressures. Collectively, these measures provide a comprehensive framework for promoting sustainable freshwater management both within the EU and in broader international contexts.

REFERENCES

- [1] J. J. Bogardi., J. Leentvaar and Z. Sebesvári, "Biologia Futura: integrating freshwater ecosystem health in water resources management," *Biologia Futura*, vol. 71, pp. 337–358, Aug. 2020, doi: 10.1007/s42977-020-00031-7. Accessed: Apr. 02, 2025. [Online]. Available: <https://link.springer.com/article/10.1007/s42977-020-00031-7>.
- [2] J. I. Santos, T. Vidal, F. J. M. Gonçalves, B. B. Castro, J. L. Pereira, "Challenges to water quality assessment in Europe – Is there scope for improvement of the current Water Framework Directive bioassessment scheme in rivers?," *Ecological Indicators*, vol. 121, Article ID: 107030, Feb. 2021, doi: 10.1016/j.ecolind.2020.107030. Accessed: Apr. 02, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1470160X20309699>
- [3] "Climate Change 2021: The Physical Science Basis - Summary for Policymakers, Technical Summary, Frequently Asked Questions and Glossary - Part of the Working Group I Contribution to the 6th Assessment Report of the Intergovernmental Panel on Climate Change," Working Group I Technical Support Unit, V. Masson-Delmotte, P. Zhai et al. Eds, PCC. Accessed: Apr. 02, 2025. [Online]. Available: https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_SummaryVolume.pdf



- [4] R. B. Bard, A. D. Eaton and E. W. Rice, Eds. *Standard Methods for the Examination of Water and Wastewater*. 23rd edition, American Public Health Association, American Water Works Association and Water Environment Federation, 2017.
- [5] J. Malinauskaite, B. Delpéch, L. Montorsi, M. Venturelli, W. Gernjak, M. Abily, T. Stepišnik Perdih, E. Nyktari and H. Jouhara, "Wastewater Reuse in the EU and Southern European Countries: Policies, Barriers and Good Practices," *Sustainability*, vol. 16, no. 24, Article ID: 11277, Dec. 2024, doi: 10.3390/su162411277. Accessed: Apr. 02, 2025. [Online]. Available: <https://www.mdpi.com/2071-1050/16/24/11277>
- [6] M. L. Partyka, R. F. Bond, "Wastewater reuse for irrigation of produce: A review of research, regulations, and risks," *Sci. Total Environ.*, vol. 828, Article ID: 154385, July 2022, doi: 10.1016/j.scitotenv.2022.154385. Accessed: Apr. 02, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0048969722014784>
- [7] W. Brack, V. Dulio, M. Ågerstrand, I. Allan, R. Altenburger, M. Brinkmann, D. Bunke, R. M. Burgess, I. Cousins, B. I. Escher, F. J. Hernández, L. Mark Hewitt, K. Hilscherová, J. Hollender, H. Hollert, R. Kase, B. Klauer, C. Lindim, D. López Herráez, C. Miège, J. Munthe, S. O'Toole, L. Posthuma, H. Rüdél, R. B. Schäfer, M. Sengl, F. Smedes, D. van de Meent, P. J. van den Brink, J. van Gils, A. P. van Wezel, A. Dick Vethaak, E. Vermeirssen, P. C. von der Ohe, B. Vrana, "Towards the review of the European Union Water Framework Directive: Recommendations for more efficient assessment and management of chemical contamination in European surface water resources," *Sci. of The Total Environment*, vol. 576, pp. 720-737, Jan. 2017, doi: 10.1016/j.scitotenv.2016.10.104. Accessed: Apr. 03, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0048969716322860>
- [8] H.E. Muga and J.R. Mihelcic, "Sustainability of wastewater treatment technologies," *J. Environ. Manage.*, vol. 88, iss. 3, pp. 437-447, Aug. 2008, doi: 10.1016/j.jenvman.2007.03.008. Accessed: Apr. 03, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0301479707001028>
- [9] G. Tchobanoglous, H. Stensel, R. Tsuchihashi and F. Burton, *Wastewater Engineering: Treatment and Resource Recovery*, 5th ed., Metcalf & Eddy, Inc. and AECOM, 2014.
- [10] D. Dutta, S. Arya and S. Kumar, "Industrial wastewater treatment: Current trends, bottlenecks, and best practices," *Chemosphere*, vol. 285, Article ID: 131245. Dec. 2021, doi: 10.1016/j.chemosphere.2021.131245. Accessed: Apr. 03, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0045653521017173>
- [11] G. Free, S. Poikane, A. Lyche Solheim, M. Bussetini, C. Bradley, J. Smith, R. Caroni, M. Bresciani, M. Pinardi, C. Giardino, W. van de Bund, "Climate change and ecological assessment in Europe under the WFD – Hitting moving targets with shifting baselines?," *J. Environ. Manage.*, vol. 370, Article ID: 122884, 2024, ISSN 0301-4797, doi: 10.1016/j.jenvman.2024.122884. Accessed: Apr. 03, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0301479724028706>.
- [12] 2023. "Waterbase - Water Quality ICM, 2022," distributed by European Environment Agency, Accessed: Apr. 03, 2025. [Online]. Available: <https://sdi.eea.europa.eu/catalogue/srv/api/records/bdead-ea2-cfaf-4724-b002-816d71c7e361>.
- [13] 2024. "Water Statistics," distributed by Eurostat, Accessed: Apr. 03, 2025. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Water_statistics
- [14] European Commission. COM/2022/541 final. (2022.10.26.). *Document 52022PC0541. Proposal for a Directive of the European Parliament and of the Council concerning urban wastewater treatment (recast)*. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0541>
- [15] "Water Quality in 2022 - An Indicators Report," Environmental Protection Agency, Ireland, Accessed: Apr. 03, 2025. [Online]. Available: <https://www.epa.ie/publications/monitoring--assessment/freshwater--marine/Water-Quality-2022-Indicator-Report-Web.pdf>
- [16] 2025. "Waste generation in Europe (Indicator)," distributed by European Environment Agency, Accessed: Apr. 03, 2025. [Online]. Available: <https://www.eea.europa.eu/en/european-zero-pollution-dashboards/indicators/waste-generation-in-europe-indicator>
- [17] F. Bichai, A. Kajenthira and S. Murthy, "Addressing barriers in the water-recycling innovation system to reach water security in arid countries," *J. Clean. Prod.*, vol.171, pp. S97-S109, Jan. 2018. Accessed: Apr. 03, 2025. [Online]. Abstract available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3147940
- [18] 2024. "SDG 6 - Clean water and sanitation," distributed by European Commission, Accessed: Mar. 23, 2025. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=SDG_6_-_Clean_water_and_sanitation
- [19] B. Grizzetti, F. Bouraoui, G. Billen, H. van Grinsven, A. C. Cardoso, V. Thieu, J. Garnier, C. Curtis, R. W. Howarth, and P. Johnes, "Nitrogen as a threat to European water quality," In *European Nitrogen Assessment*, M. A. Sutton, C. M. Howard, J. W. Erismann, G. Billen, A. Bleeker, P. Grennfelt, H. van



Grinsven, and B. Grizzetti, Eds., Cambridge and Reading, United Kingdom: Cambridge University Press and University of Reading, 2011, ch. 17, pp. 379-404. [Online]. Available: https://centaur.reading.ac.uk/20869/1/28387ENA_c17.pdf

- [20] R. A. A. Meena, R. Yukesh Kannah, J. Sindhu, J. Ragavi, G. Kumar, M. Gunasekaran, and J. Rajesh Banu, "Trends and resource recovery in biological wastewater treatment system," *Bioresour Technol. Rep.*, vol. 7, Article ID: 100235, Sep. 2019, doi: 10.1016/j.biteb.2019.100235. Accessed: Apr. 03, 2025. [Online]. Available: <https://yonsei.elsevier-pure.com/en/publications/trends-and-resource-recovery-in-biological-wastewater-treatment-s>



THE ROLE OF OPEN INNOVATION IN ENHANCING MANAGERIAL COMPETENCIES IN AI-DRIVEN SYSTEMS

Tamara Papić^{1*},
[0000-0002-9666-7824]

Bojan Gutić¹,
[0000-0002-1085-4718]

Nina Pantelić¹,
[0009-0005-3401-8537]

Nikola Petrović²
[0009-0001-7750-4712]

¹Singidunum University,
Belgrade, Serbia

²Department of Industrial Engineering
and Management,
Faculty of Technical Science,
University of Novi Sad,
Novi Sad, Serbia

Abstract:

As businesses undergo digital transformation, artificial intelligence (AI) is becoming an integral part of decision-making and operational processes. However, managing AI-driven systems effectively requires a multidisciplinary approach that blends technical expertise, strategic foresight, and adaptive management skills.

This paper examines how open innovation contributes to the development and refinement of managerial competencies in organizations that incorporate AI technologies and the relationship between AI and open innovation, focusing on how the AI and open innovation concepts enhance decision-making and collaboration in business environments. By reviewing relevant literature and industry cases, we provide insights into AI's role in modern management and suggest directions for future research.

The findings indicate that organizations leveraging open innovation strategies—such as collaborative research and development, external partnerships, and crowdsourced problem-solving—demonstrate greater adaptability and efficiency in AI-driven management. Furthermore, this paper proposes a framework for integrating AI capabilities with business management practices, emphasizing the role of dynamic leadership and innovation ecosystems. By offering insights into how firms can optimize their AI strategies through open innovation, this study contributes to both academic research and practical applications.

The results suggest that organizations adopting these approaches can enhance performance, strengthen their competitive edge, and drive sustainable growth.

Keywords:

Open Innovation, AI-Driven Management, Business Competencies, Knowledge Sharing, Digital Transformation.

INTRODUCTION

Artificial intelligence (AI) is reshaping industries by automating processes, making systems more scalable, enhancing decision-making, and driving innovation. AI systems usually leverage machine learning, predictive analytics, natural language processing, and data visualization, with the aim of analyzing huge amounts of data, enabling businesses to optimize workflows, improve customer experiences, and gain a competitive edge [1] [2]. The adoption of AI extends across multiple sectors, including services, manufacturing, finance, healthcare, media, marketing and is used for sales channel support, CRM, fraud detection, medical diagnosis, supply chain optimization, and personalized communication and recommendations [3] [4].

Correspondence:

Tamara Papić

e-mail:

tpapic@singidunum.ac.rs





Beyond its technological capabilities, AI demands a fundamental shift in business strategy and management practices. Successful implementation requires more than just advanced algorithms; it necessitates an open innovation approach that fosters collaboration between businesses, research institutions, and technology developers [5] [6]. By leveraging external knowledge and interdisciplinary expertise, organizations can accelerate AI-driven advancements, enhance adaptability, and create sustainable competitive advantages in an increasingly digital economy.

Rapid technological advances are reshaping industries, with many organizations increasingly adopting artificial intelligence (AI) to enhance their open innovation strategies. Open innovation, a concept introduced by Chesbrough [1], promotes the integration of internal and external sources of knowledge to drive innovation and maintain a competitive advantage. Unlike the traditional closed innovation model, where research and development (R&D) activities remain confined within the company, open innovation encourages collaboration with external partners, such as research institutions, startups, technology firms, and industry experts. As technological advances become more complex and market dynamics become increasingly unpredictable, companies embrace this approach to improve adaptability, accelerate innovation cycles, and create value through shared expertise and collaborative networks. From a management perspective, AI plays a key role in powering open innovation by simplifying data analysis, optimizing cross-organizational collaboration, and supporting informed decision-making. Studies show that companies incorporating AI into their open innovation frameworks experience accelerated product development, improved problem-solving capabilities, and greater responsiveness to market fluctuations [6] [7]. Machine learning and AI-driven data analytics enable organizations to process vast amounts of previously underutilized data, extracting valuable insights that align innovation strategies with evolving consumer preferences and industry trends [8] [9]. Moreover, AI tools facilitate seamless collaboration with external stakeholders, including universities, research centers, technology hubs, and innovation consultants—by identifying synergies and optimizing the exchange of knowledge and expertise.

In addition to improving internal efficiency, AI provides managers with deeper insights into consumer behavior, market trends, and the competitive landscape. AI-based decision support systems enable data-driven strategic planning, reducing reliance on intuition, and enabling businesses to proactively adapt to internal and

external changes [10] [11]. By integrating AI into open innovation frameworks, organizations can develop more agile business models that promote continuous learning, rapid prototyping, and collaborative problem-solving [12] [13]. In addition, AI improves knowledge transfer within and across corporate boundaries by automating data processing, improving information accessibility, and fostering interdisciplinary communication [14].

However, integrating AI into open innovation governance poses significant challenges. Ethical concerns, such as algorithmic biases, data privacy risks, and the transparency of AI-driven decisions, have sparked discussions about accountability and fairness in AI applications [15] [16]. Furthermore, organizational resistance, including “algorithm aversion”—where employees and decision makers are skeptical of AI-generated insights, despite their proven accuracy—can hinder the adoption of AI in critical business functions [12]. Addressing these challenges requires the establishment of ethical AI frameworks, transparent and explainable AI systems, and targeted initiatives to build trust in AI-driven governance processes [17].

As AI technologies continue to evolve, their impact on business growth, innovation ecosystems, and governance structures will become increasingly significant. Organizations that embrace open innovation and strategically integrate AI into their decision-making processes will be better positioned to maintain their competitive advantage in dynamic markets [18]. Finally, this paper highlights the importance of balancing AI-driven automation with human expertise, ensuring that AI serves as a complementary tool that enhances governance efficiency, rather than replacing human judgment [19].

2. METHODOLOGY AND RESULTS

This research applies a Systematic Literature Review (SLR) and empirical analysis to investigate the role of open innovation in the adoption of AI in business management. The methodology consists of two key stages:

Systematic Literature Review (SLR): A comprehensive analysis of peer-reviewed journal and conference articles from IEEE Xplore, Web of Science, and Scopus, from the last five years (2019–2025). The selection criteria focus on:

- Integration of AI in decision-making and business management
- Open innovation strategies for implementing AI
- Case studies of companies successfully adopting artificial intelligence through open innovation



2.1. OPEN INNOVATION CONTRIBUTION TO MANAGEMENT

Open innovation helps businesses stay agile, collaborate effectively, and adopt modern technologies faster. By working with startups, universities, and research institutions, companies can foster innovation while reducing the risk of developing everything in-house. As shown in Table 1, for management, this approach supports flexible decision-making, access to advanced technologies, and better resource allocation. AI-powered open innovation further improves market analysis, trend identification, and strategic planning. The next section examines its impact on various industries and how it strengthens decision-making [1] [2] [3].

2.2. INDUSTRIES ENGAGED IN OPEN INNOVATION

The adoption of open innovation varies across industries, which are influenced by market dynamics, technological advancements, and competitive pressure [1]. Figure 1 demonstrates the percentage of companies

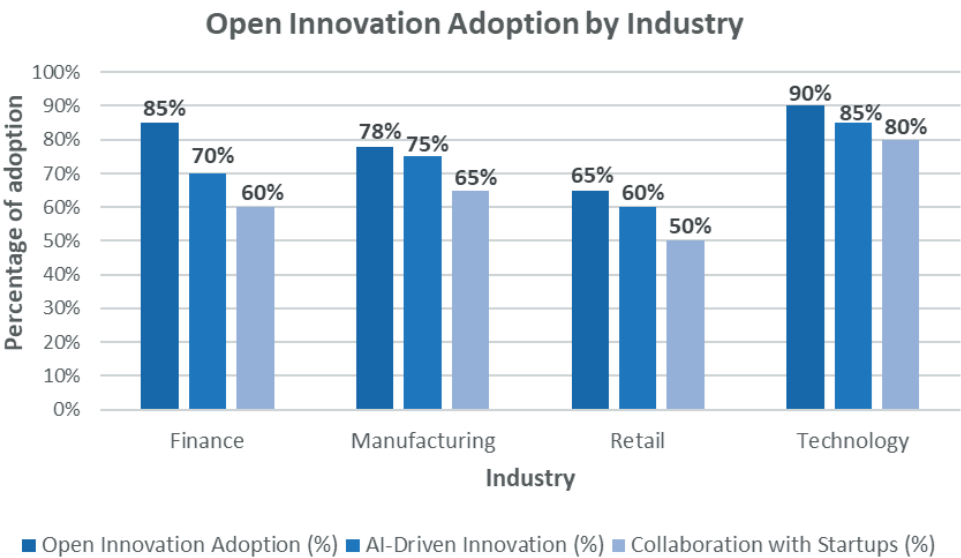
across industries that have adopted open innovation strategies, implemented AI-led innovation, and collaborated with startups. As shown, the technology sector leads the way in adopting open innovation (90%), followed by finance (85%), manufacturing (78%), and retail (65%). AI-led innovation is closely aligned with open innovation adoption, indicating a strong correlation between these two transformative approaches [2].

2.3. AI USAGE IN BUSINESS MANAGEMENT FUNCTIONS

AI technologies are increasingly being applied across various management functions to improve business performance. Figure 2 demonstrates the distribution of AI usage across four key business areas: data analytics, process automation, user engagement, and strategic decision-making [20]. The highest adoption was observed in data analytics (30%), followed by process automation (25%) and strategic decision-making (25%). These trends suggest that businesses are prioritizing AI-driven insights and automation to maintain a competitive advantage in their industries [21].

Table 1. Key Areas of Open Innovation Contribution to Business Management

Open Innovation Contribution	Impact on Management
External collaboration	Faster technological adoption, access to expertise
Ecosystem partnerships	Strengthened R&D and co-innovation potential
Data sharing & AI training	Improved decision-making capabilities
Reduced innovation costs	More efficient resource allocation
Business model transformation	Enhanced competitive advantage



Listing 1. Open Innovation Adoption by Industry



2.4. AI IMPLEMENTATION BENEFITS FOR BUSINESS DECISION-MAKING

Artificial intelligence plays a pivotal role in enhancing business decision-making processes, optimizing operations, and providing strategic advantages. Figure 3 highlights how different AI applications contribute to improved decision-making efficiency, process optimization, and competitive advantage. Predictive decision-making demonstrates the highest impact (95%), followed by data analytics (80%), and process automation (85%). These findings suggest that organizations leveraging AI can achieve significant gains in efficiency and strategic planning.

2.5. THE CASE STUDY OF BMW AND UNILEVER

BMW was one of the first companies to adopt artificial intelligence through open innovation with the aim of improving its manufacturing processes and product development. The company partnered with several external startups and research institutions in order to integrate AI solutions into its business. A key area of focus for BMW is improving production efficiency and the overall user experience, especially with the context of autonomous vehicles. BMW's collaboration with the University of California, Berkeley, and AI startups such as aiMotive, facilitated their leverage of AI into real-world applications, such as vehicle safety features, predictive maintenance, and autonomous driving technologies [22] [23].

Percentage of AI Usage in Business Management Functions

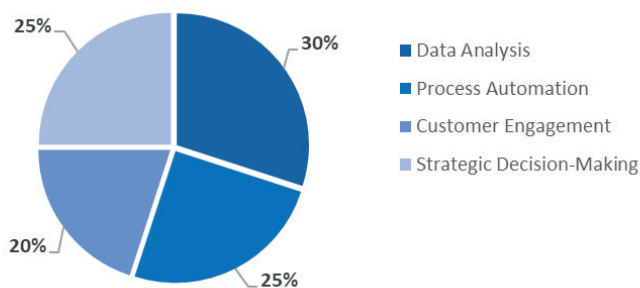


Figure 2. Percentage of AI Usage in Business Management Functions

AI Implementation Benefits for Business Decision-Making

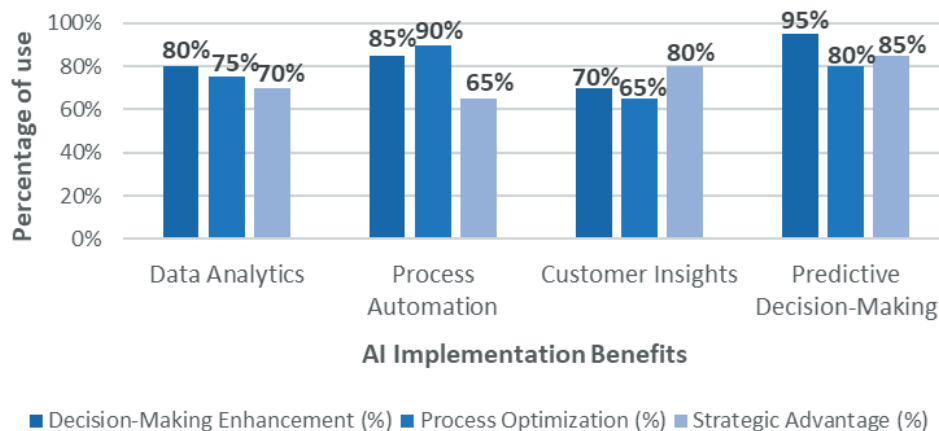


Figure 3. AI Implementation Benefits for Business Decision-Making



Unilever embraced artificial intelligence and digital technologies to enhance its product development processes. The company is working with various partners to leverage data and artificial intelligence with the goal of accelerating innovation and meeting evolving consumer needs. Unilever's approach involves using artificial intelligence to analyze consumer behavior, predict market trends, and design products that resonate with consumers around the world. This open innovation strategy allows Unilever to integrate external insights and technologies into its R&D efforts, thereby fostering a culture of continuous improvement and responsiveness to market demands [24] [25].

3. DISCUSSION

The integration of artificial intelligence supported by open innovation-driven ecosystems has fundamentally changed the way companies collaborate and innovate. Open innovation, as defined by Chesbrough [1], emphasizes the importance of leveraging external knowledge for internal needs and innovation. Here, AI plays a key role by improving decision-making and fostering collaboration across organizations and industries.

3.1. OPEN INNOVATION ECOSYSTEMS AND ARTIFICIAL INTELLIGENCE

Aldoseri et al. [9] state that artificial intelligence plays a fundamental role in creating dynamic and adaptive open innovation systems, enabling managers to discover and implement new business models and opportunities. The participation of AI in these systems helps managers accelerate the validation of ideas, thus improving the possibilities of companies responding quickly to changing and emerging markets and evolving customer preferences.

3.2. STRATEGIC DECISION-MAKING

The impact of artificial intelligence is significantly expanding to strategic decision-making. T. Broekhuizen et al. [4] emphasize that AI technologies can support managers and organizations in making more informed decisions through the analysis of large data sets. This capability allows companies to proactively adapt to different market challenges, change existing business models, and anticipate future market challenges [6].

3.3. AI AND BUSINESS MODEL INNOVATION

AI's potential to drive innovation in business models is substantial, as outlined by R. C. Climent et al. [13]. They argue that AI enables firms to embrace more agile and flexible business models, particularly when integrated with the principles of open innovation. By leveraging AI, organizations can gain deeper insights into market dynamics and consumer behavior, which fosters the development of innovative and sustainable business models [26] [27].

3.4. CHALLENGES AND OBSTACLES

While the integration of AI into open innovation frameworks offers numerous advantages, it also presents significant challenges. A. Kuzior et al. [12] note that companies face hurdles, such as data security, system interoperability, and resistance to technological change when adopting AI. Additionally, A. Taleghani et al. highlights the importance of cultivating the requisite skills and organizational capabilities necessary for fully harnessing AI's potential in open innovation ecosystems [21].

4. CONCLUSION

The role of AI in transforming open innovation is evident across industries, from improving decision-making to enabling more dynamic and flexible business models. However, as companies continue to adopt AI into their innovation processes, it is crucial to address the associated challenges to fully unlock its potential. As companies move forward, the integration of AI into open innovation ecosystems will play a key role in shaping future competitive landscapes.



REFERENCES

- [1] H. Chesbrough, *Open Innovation: The New Imperative for Creating and Profiting from Technology*, Harvard Business Press, 2003.
- [2] J. West and M. Bogers, "Open Innovation: A Research Agenda," *Journal of Product Innovation Management*, vol. 31, no. 6, pp. 1199-1217, 2014.
- [3] G. Secundo, C. Spilotro, J. Gast and V. Corvello, "The transformative power of artificial intelligence within innovation ecosystems: a review and a conceptual framework," *Review of Managerial Science*, vol. 18, pp. 1-32, 2024.
- [4] T. Broekhuizen, H. Dekker, P. d. Faria, S. Firk, D. K. Nguyen and W. Sofka, "AI for managing open innovation: Opportunities, challenges, and a research agenda," *Journal of Business Research*, vol. 167, 2023.
- [5] D. J. Teece, "Business Models, Business Strategy, and Innovation," *Long Range Planning*, vol. 43, no. 2, pp. 172-194, 2010.
- [6] A. Abositta, M. W. Adedokun and A. Berberoğlu, "Influence of Artificial Intelligence on Engineering Management Decision-Making with Mediating Role of Transformational Leadership," *Systems*, vol. 12, no. 12, p. 570, 2024.
- [7] M. L. Prasetyo, R. A. Peranginangin, N. Martinovic, M. Ichsan and H. Wicaksono, "Artificial intelligence in open innovation project management: A systematic literature review on technologies, applications, and integration requirements," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 11, no. 1, p. 100445, 2025.
- [8] D. L. Roberts and M. Candi, "Artificial intelligence and innovation management: Charting the evolving landscape," *Technovation*, vol. 136, 2024.
- [9] A. Aldoseri, K. N. Al-Khalifa and A. M. Hamouda, "AI-Powered Innovation in Digital Transformation: Key Pillars and Industry Impact," *Sustainability*, vol. 16, no. 5, 2024.
- [10] J. C. C. Sin and V. Kathiarayan, "The Role of Artificial Intelligence in Strategic Decision-Making: Opportunities, Challenges, and Implications for Managers in the Digital Age," *International Journal of Management and Commerce Innovations*, vol. 11, no. 1, pp. 73-79, 2023.
- [11] M. Holgersson, L. Dahlander, H. Chesbrough and M. L. A. M. Bogers, "Open Innovation in the Age of AI," *California Management Review*, vol. 67, no. 1, pp. 5-20, 2024.
- [12] A. Kuzior, M. Sira and P. Brożek, "Use of artificial intelligence in terms of open innovation process and management," *Sustainability*, vol. 15, no. 9, p. 7205, 2023.
- [13] R. C. Climent, D. M. Haftor and M. W. Staniewski, "AI-enabled business models for competitive advantage," *Journal of Innovation & Knowledge*, vol. 9, no. 3, 2024.
- [14] V. Jain, K. Wadhvani and J. K. Eastman, "Artificial intelligence consumer behavior: A hybrid review and research agenda," *Journal of Consumer Behaviour*, vol. 23, no. 3, 2023.
- [15] B. Y. Kassa and E. K. Worku, "The impact of artificial intelligence on organizational performance: The mediating role of employee productivity," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 11, no. 1, p. 100474, 2025.
- [16] P. Jorzik, S. P. Klein, D. K. Kanbach and S. Kraus, "AI-driven business model innovation: A systematic review and research agenda," *Journal of Business Research*, vol. 172, 2024.
- [17] S. Kaggwa, T. F. Eleogu, F. O. Okonkwo and O. Farayola, "AI in decision making: Transforming business strategies," *Int. J. Res. Sci. Innov*, vol. X(XII), pp. 423-444, 2024.
- [18] I. A. M. FuelCell Energy, "Predictive analytics in business: Leveraging AI for competitive advantage," *IPE Journal of Management*, 2025.
- [19] B. Andres, D. Fernandez-Garcia, P. Urze and E. Araujo, "Artificial Intelligence use in collaborative network processes," 2024.
- [20] Qiu and B. Chang, "The impact of digital transformation on open innovation performance: The intermediary role of digital innovation dynamic capability," *PLOS ONE*, vol. 20, no. 3, 2025.
- [21] A. Taleghani, M. Taleghani and S. J. Mousavian, *The Place and Role of Artificial Intelligence in the Open Innovation of Start-Ups (Opportunities and Challenges)*, 2023.
- [22] BMW Group, "BMW Open Innovation," BMW Group. [Online].
- [23] BMW Group, "Artificial Intelligence," BMW Group. [Online].
- [24] E. Enkel, O. Gassmann and H. W. Chesbrough, "Open R&D and Open Innovation: Exploring the Phenomenon," *R&D Management*, vol. 39, no. 4, pp. 311-316, 2009.
- [25] Unilever, "How AI and digital help us innovate faster and smarter," Unilever News, April 2023. [Online].
- [26] D. J. Teece, "Business models and dynamic capabilities," *Long Range Planning*, vol. 51, no. 1, pp. 40-49, 2018.
- [27] D. Sjödin, V. Parida, M. Palmié and J. Wincent, "How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops," *Journal of Business Research*, vol. 143, pp. 456-469, 2022.



INFORMATION TECHNOLOGY IN TEACHING FOREIGN LANGUAGES SESSION





ARTIFICIAL INTELLIGENCE IN MOBILE LANGUAGE LEARNING: DUOLINGO AND THE RISE OF A NEW EDUCATIONAL ERA

Tijana Gajić*,
[0000-0002-5013-2867]

Jelena Nikolić,
[0000-0002-7013-9513]

Neda Maenza,
[0000-0002-4974-6052]

Aleksandra Gajić
[0000-0002-6804-4662]

Singidunum University,
Belgrade, Serbia

Abstract:

This paper investigates the role of artificial intelligence (AI) in mobile applications for language learning, with a particular focus on Duolingo. It analyses how AI enhances personalized learning through adaptive technologies and features such as speech recognition and interactive exercises. Additionally, the paper explores AI's advantages, such as automated grading and reduced administrative tasks for teachers. Special emphasis is placed on how these technologies are transforming the educational process, considering their impact on both teachers and students. The discussion also addresses potential limitations, including concerns over data privacy and the over-reliance on technology in education. The conclusion acknowledges that while AI tools can significantly enhance the learning process, they cannot fully replace the human factor, especially in the domains of emotional intelligence and critical thinking. Ultimately, the paper argues that AI should be seen as a complementary tool rather than a substitute for traditional teaching methods, fostering a more balanced and effective approach to language learning.

Keywords:

Artificial Intelligence, Mobile Applications, Duolingo, Language Learning, Personalized Learning.

INTRODUCTION

We are witnessing the profound impact of artificial intelligence on foreign language learning [1]. In recent years, AI-based mobile language learning applications have become widely adopted tools in both formal and informal education. These applications leverage advanced algorithms, natural language processing, and adaptive technologies to provide personalized learning, customizing content to each user's needs, offering real-time feedback, and simulating interactive communication in the target language [2] [3]. This raises the question of whether traditional language learning approaches are facing challenges while, at the same time, a new era of education is emerging in which AI is redefining how we approach foreign language acquisition. However, concerns about data privacy and the ethical aspects of artificial intelligence continue to trouble many users. This paper investigates the role of artificial intelligence in mobile language learning applications and their potential impact on the future of education, focusing on new AI-based features of one of the world's most popular language-learning applications, Duolingo.

Correspondence:

Tijana Gajić

e-mail:

tgajic@singidunum.ac.rs





2. THE CHALLENGE OF ARTIFICIAL INTELLIGENCE IN FOREIGN LANGUAGE LEARNING FOR TEACHERS

With AI's profound influence on all spheres of education, including foreign language acquisition, the role of teachers in the classroom is inevitably changing. This raises the question of whether AI is a threat to language teachers or a valuable ally in the complex process of language acquisition. The authors of this paper, initially sceptical, have come to adopt a positive view of AI in education. We cannot ignore its presence or the fact that our students recognize its advantages as well as its disadvantages. AI is here to stay. Teachers must adapt to this new reality and help students use AI tools appropriately. Moreover, AI-based applications provide diverse resources that can enrich traditional curricula. AI is a valuable ally, particularly for teachers managing large student groups, as applications for automated grading accurately assess each student's progress and are often used for formative assessment. AI can also significantly reduce teachers' administrative burdens. In recent years, educators have increasingly used AI tools like Fetchy, a virtual assistant for teachers. Fetchy can be utilized in numerous ways to simplify and streamline teachers' professional responsibilities, offering inspiration and personalization for various teaching scenarios [4].

One of the authors' key recommendations is to use classrooms primarily for discussions, case studies, practical activities, and other tasks aimed at developing critical thinking, thus preserving them as spaces for social emancipation in line with key concepts of critical pedagogy [5]. On the other hand, AI applications designed to support autonomous learning can be used at home, aligning with the flipped classroom model.

Research shows that AI tools yield the best results when they complement, rather than replace, traditional teaching methods and that a combination of both approaches leads to better academic performance and greater student engagement [6] [7] [8].

Using AI grants teachers complete autonomy in their teaching processes, enabling them to monitor students' progress and adjust teaching materials accordingly. AI allows teachers to create additional exercises for areas where students struggle or need to review. As a result, the language learning process is optimized, with more time spent on difficult and complex areas and less on those students have already mastered. Creating supplementary exercises is highly efficient since teachers can generate a variety of exercises targeting specific linguistic challenges quickly.

Notably, AI offers significant benefits in ESP (English for Specific Purposes), particularly in creating vocabulary exercises, given the difficulty of finding textbooks and materials that are comprehensive and aligned with the rapidly changing market demands, as well as the needs of professors and students adapting to these changes. By simply inputting the target vocabulary and the desired language level, teachers can generate a ready-made text with various vocabulary exercises in just a few minutes.

The modifications AI makes are particularly important for students with disabilities, as AI tools can help create accessible learning materials tailored to individual needs, such as text-to-speech features, adaptive exercises, and personalized learning paths.



Figure 1. Fetchy virtual assistant for educators



Given that AI is an inexhaustible source of information that can be consulted without the risk of being too demanding, anything undesirable can be modified and adapted to suit both teachers' and students' needs. Refreshing outdated materials is especially engaging, as it introduces greater dynamics into lessons and breaks the monotony that inevitably transfers from professor to students when the same materials are used year after year.

In addition to all the advantages of AI tools in education, we must also address the main drawbacks. These include, primarily, the issue of data privacy and the ethical aspects of artificial intelligence, as well as the time and effort teachers invest in their professional development to feel confident when using modern technologies in teaching. If an application shares user information with marketing firms, users might remain unaware that their data is being disclosed. Applications frequently use gamification to engage learners. However, excessive gamification can create dependency, causing users to feel pressured or anxious about daily app use.

For example, an article in *El País* highlights how Duolingo's mascot, Duo the Owl, employs persistent and emotionally manipulative tactics to keep users engaged, inducing guilt or stress in some learners [9].

In the context of the drawbacks of AI, the authors must also address the business policies of Duolingo's founder and CEO, Luis von Ahn. At the end of 2023, Duolingo decided not to renew contracts with around 10% of its freelance workers, replacing them with AI-generated translations and lessons in some cases. "Our stance as a company is that if we can automate something, we will," said founder and CEO Luis von Ahn regarding the layoffs [10]. In the long term, von Ahn optimistically adds, AI could open up new opportunities for gaining knowledge, providing quality education to the masses. He believes that languages, in particular, can help people overcome poverty.

3. DUOLINGO – AN AI-BASED LANGUAGE LEARNING APPLICATION

Duolingo is a pioneer in AI innovations for language learning. AI algorithms meticulously analyse user performance, continuously adjusting lesson difficulty to match individual progress. This personalized aspect ensures that Duolingo lessons are equally challenging and appropriate for learners at all levels.

Duolingo has recently introduced an interactive feature where users engage in video calls with Lily, one of Duolingo's beloved mascots. She is a slightly sarcastic girl with purple hair. Conversing with Lily allows users to practice speaking in other languages as if they were talking to an AI friend. These dialogues are generated using the OpenAI GPT-4 model. The feature is part of a subscription package called Duolingo Max, which costs \$30 per month and was launched in 2023 for its premium AI functionalities [11].

Premium AI features include the 'Explain My Answer' option, which provides detailed and personalized explanations for all incorrect answers. One of the biggest criticisms of the standard free Duolingo app is the lack of explanations when users make mistakes. The app would alert learners if they were incorrect and show the correct answer. Duolingo Max, powered by AI, strives to give all its users an insight into the principles governing a language's structure.

Another new AI addition is the "Roleplay" option, which places users in interactive real-life situations to practice language skills, such as ordering coffee, checking a passport, or asking for directions. Many studies have shown that speaking skills can be significantly improved through the regular use of AI applications [12] [13]. For many users, the novelty of practicing with AI characters could provide additional motivation to continue learning the language.



Figure 2. Duolingo mascot Lily



Table 1 clearly shows that the user dropout rate has decreased by as much as 25%, and in terms of results, users are 30% more successful compared to the period before AI implementation.

However, judging by the feedback from early users, while Roleplay is a remarkably helpful feature for practicing conversation, AI models have limitations. Their responses still cannot replace conversations with native speakers [14].

These features represent the latest wave of tools from the generative AI movement initiated by the company last year. “In my opinion, a personalized AI tutor is not a particular feature we are building,” said Klinton Bicknell, head of the AI team at Duolingo, in an interview with Forbes. “It is more of a vision of what the entire app is becoming” [10].

The AI approach implemented in the Duolingo app has resulted in an increase in users and revenue after going public in 2022 [10]. Almost 104 million people use the app to learn languages, math, and music each month, representing a 40% increase compared to the previous year. This quarter, revenue has amounted to \$178.3 million, a 41% increase compared to last year [11].

In addition to efforts to develop an AI tutor, the company has made another significant investment in AI through its Duolingo English Test (DET). This is an application-based version of the TOEFL test used worldwide to certify English language proficiency for university or visa applications [11].

The DET, which costs \$59 and was first launched in 2016, became popular during the COVID-19 pandemic because it could be taken remotely. Duolingo now uses AI for every aspect of the DET exam (Forbes), from generating questions to ensuring test-takers do not cheat. One security feature, for example, uses facial recognition to ensure that the examinee is not looking at notes off-screen. Currently, this test accounts for 10% of Duolingo's revenue, and the founder and CEO, Luis von Ahn, aims to further increase revenue, as the company is increasingly targeting non-native English speakers.

Elizabeth Birr Moje, dean of the Marsal Family School of Education at the University of Michigan, said she was excited about the potential of Duolingo's new AI features. However, she does not believe that AI tools will ever be able to replace the intangible skills of real teachers. "AI can't see if a student is experiencing frustration. It cannot see body language," she said. "It can't see joy" [10].

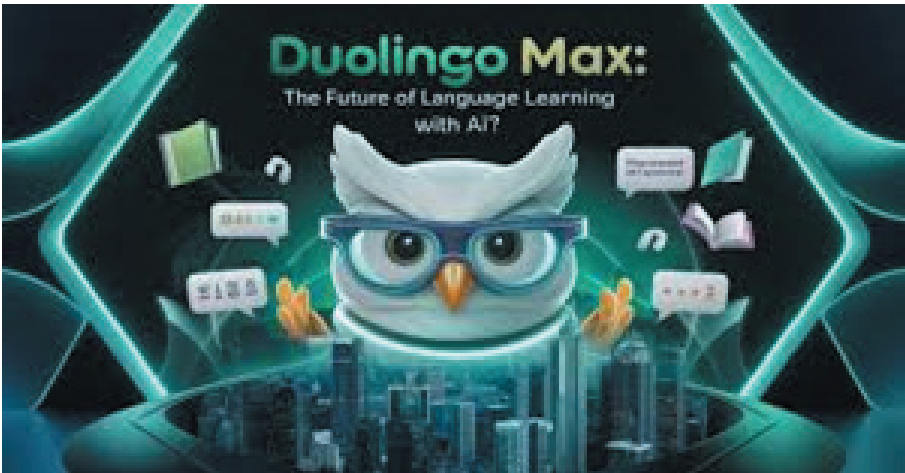


Figure 3. The Future of Language Learning with AI

Table 1. Implementation Statistics and Cost Information

Implementation Statistics and Cost Information
Learning Outcomes - 30% improvement in proficiency levels.
Implementation Cost - \$8 million for adaptive learning AI integration.
ROI - Realized within 15 months with increased user satisfaction.
User Retention - 25% reduction in dropout rates.



AI tools like Duolingo are primarily designed for individual learning, allowing users to progress at their own pace and focus on their specific needs. However, there is significant potential for integrating collaborative learning into these platforms, where students could work together or in small groups with the support of artificial intelligence [15]. Through AI technologies, applications could enable interactive tasks that require collaboration, such as joint problem-solving, idea-sharing, or multi-participant conversation simulations [16] [17]. This would not only enrich the learning experience through social interaction but also foster the development of communication skills in a more realistic context.

To conclude, Duolingo's development reflects larger trends in technology-based language education, highlighting the crucial role of innovation in addressing the changing needs of learners in the digital era [18].

In addition to the Duolingo app, there are apps like Babbel, Pronounce, LanguaTalk, and TalkPal that also use artificial intelligence to enhance the educational experience in numerous ways. These technologies also support autonomous learning and inclusivity, allowing users to learn languages at their own pace.

4. CONCLUSION

The role of artificial intelligence in mobile language learning applications has become a crucial factor in modern education, given its potential to enhance personalized learning and interactive communication. These technologies provide tailored content and instant feedback and foster independent learning. Although artificial intelligence brings many advantages, such as facilitating the teaching process, optimizing formative assessment, and reducing teachers' administrative workload, the obstacles it introduces should not be overlooked. First and foremost, there are ethical concerns related to data privacy and the potential impact on employment opportunities in the education sector. Ethical responsibility is a key component in creating mobile applications based on artificial intelligence. The example of Duolingo, which uses AI to perform tasks that were once handled by humans, illustrates how this technology is transforming the job market and the educational landscape. Despite these challenges, research shows that combining traditional learning methods with AI tools is the most effective way to achieve better results and engage students. In this regard, AI can be a valuable ally to teachers. However, it cannot fully replace the human

element in education, especially in areas that require emotional intelligence, critical thinking, and social interaction. Moving forward, ongoing collaboration between educators, developers, and policymakers will be essential to harness the full potential of AI while maintaining pedagogical integrity. In the long run, the successful integration of artificial intelligence in education will depend on a careful balance between technological innovations and the preservation of key pedagogical values, ensuring that AI serves as an enhancement rather than a substitute for meaningful human interaction in the learning process.

REFERENCES

- [1] X. Huang, D. Zou, G. Cheng, X. Chen and H. Xie, "Trends, Research Issues and Applications of Artificial Intelligence in Language Education," *Educational Technology & Society*, vol. 26, no. 1, pp. 112-131, 2023.
- [2] B. Zou, X. Guan, Y. Shao and P. Chen, "Supporting Speaking Practice by Social Network-Based Interaction in Artificial Intelligence (AI)-Assisted Language Learning," *Sustainability*, vol. 15, no. 4, 2023.
- [3] N. Hockly, "Artificial Intelligence in English Language Teaching: The Good, the Bad and the Ugly," *RELC Journal*, vol. 54, no. 2, pp. 445-451, 2023.
- [4] "Fetchy," [Online]. Available: <https://www.fetchy.com/>. [Accessed 11 January 2025].
- [5] T. Gajić and A. Kuzmanović Jovanović, *Foreign Language Teaching and the Development of Critical Thinking in the Era of Neoliberalism: A Critical Pedagogy Perspective*, Belgrade: Čigoja Press, 2022.
- [6] O. Zawacki-Richter, V. I. Marín, M. Bond and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 39, 2019.
- [7] W. Holmes, M. Bialik and C. Fadel, *Artificial Intelligence In Education: Promises and Implications for Teaching and Learning*, Boston: The Center for Curriculum Redesign, 2019.
- [8] E. Brynjolfsson and A. McAfee, "The Business of Artificial Intelligence," *Harvard Business Review*, 2017.
- [9] M. Alonso, "Hace mucho que no te vemos": cómo Duolingo hizo de la pasivo-agresividad su secreto adictivo," *El País*, 2024.
- [10] R. Nieva, "Na jednoj od omiljenih aplikacija za jezike Duolingo uskoro će nas podučavati AI profesor," *Forbes SRB*, 2024.



- [11] "Duolingo - The world's best way to learn a language," [Online]. Available: <https://www.duolingo.com/>. [Accessed 12 January 2025].
- [12] G. Dizon, "Evaluating Intelligent Personal Assistants for L2 Listening and Speaking Development," *Language Learning & Technology*, vol. 1, no. 16-26, p. 24, 2020.
- [13] K. W. K. Lai and H. J. H. Chen, "An exploratory study on the accuracy of three speech recognition software programs for young Taiwanese EFL learners," *Interactive Learning Environments*, vol. 32, no. 5, pp. 1582-1596, 2024.
- [14] E. Caldwell, "Duolingo Max: The Future of Language Learning with AI?," Duolingo Guides, 2024.
- [15] N. K. Tuong and T. C. Dan, "A Study on Duolingo Mobile Applications to Improve Efl Students' Listening Comprehension Performances," *European Journal of Alternative Education Studies*, vol. 9, no. 1, 2024.
- [16] Z. Ouyang, Y. Jiang and H. Liu, "The Effects of Duolingo, an AI-Integrated Technology, on EFL Learners' Willingness to Communicate and Engagement in Online Classes," *International Review of Research in Open and Distributed Learning*, vol. 25, no. 3, pp. 97-115, 2024.
- [17] R. K. Yekollu, T. B. Ghuge, S. S. Biradar, S. V. Haldikar, S. V. Haldikar and O. F. M. A. Kader, "AI-Driven Personalized Learning Paths: Enhancing Education Through Adaptive Systems," in *Smart Data Intelligence*, Singapore, 2024.
- [18] J. Vega, M. Rodriguez, Check, Erick, H. Moran, Loo and Luis, "Duolingo evolution: From automation to Artificial Intelligence," in *2024 IEEE Colombian Conference on Applications of Computational Intelligence (ColCACI)*, Pamplona, 2024.



THE DIGITAL DOPPELGÄNGERS OF NIKOLA TESLA AND BRANISLAV NUŠIĆ: A NEW APPROACH TO INTERACTIVE LEARNING AND CULTURAL HERITAGE

Milena Škobo^{1*},
[0000-0001-8427-2051]

Milena Šović²
[0009-0009-5774-5672]

¹Sinergija University,
Bijeljina, Republic of Srpska

²University Business Academy in
Novi Sad, Faculty of Economics and
Engineering Management,
Novi Sad, Serbia

Abstract:

With the advancement of AI and digital doppelgängers, the creation of virtual replicas of historical figures is becoming increasingly significant in the fields of culture, education, and scientific research. This paper analyzes theoretical aspects of developing digital doppelgängers and introduces a new subcategory—cultural-educational chatbots—through the examples of pioneering Serbian chatbot projects modeled after Nikola Tesla and Branislav Nušić. These pilot chatbots, currently undergoing testing in Serbian schools and museums, represent the first cultural-educational chatbots of their kind in Serbia. The study examines the technical, linguistic, and ethical challenges associated with replicating historical figures, exploring critical issues such as authenticity, cultural context, and the limits of digital immortality. Through the ongoing pilot testing, this paper highlights both the innovative potential and the practical considerations for the broader adoption of such chatbots in Serbian educational and cultural sectors.

Keywords:

AI, Digital Doppelgängers, Tesla Chatbot, Nušić Chatbot, Educational and Cultural Sector.

INTRODUCTION

In the age of advanced AI and digital doppelgängers, the attempt to “bring historical figures back to life” through chatbots is becoming an increasingly intriguing challenge. But how feasible is it, really, to replicate the mind and spirit of an individual—especially when it comes to greats like Nikola Tesla and Branislav Nušić? This raises an essential question: can we ever create an authentic replica of someone who lived more than a century ago? The paradigm of digital doppelgängers is closely tied to conversational AI. Today, different types of digital doppelgängers exist—from so-called “griefbots,” which simulate communication with deceased loved ones, to projects like DigiDan, a chatbot modeled after the philosophical style of Daniel Dennett. The development of these technologies opens new possibilities, particularly in education, but also raises significant ethical concerns.

Correspondence:

Milena Škobo

e-mail:

mskobo@sinergija.edu.ba





This paper explores the challenges of creating digital doppelgängers of historical figures through the examples of the Nikola Tesla and Branislav Nušić chatbots—the first pilot cultural-educational chatbots in Serbia. In primary and secondary education, these chatbots can enhance the learning process and make teacher-student interactions more dynamic and immersive. By “reviving” the author or the scientist and speaking in their name, they can make learning more accessible and enjoyable while inspiring more engaging and motivating teaching experiences. In cultural institutions such as national museums, galleries, and endowments, these tools can promote cultural heritage in innovative and modern ways, fostering a stronger bond between visitors through interactive chatbot experiences. However, their potential development and usage brings up critical questions, including the selection of historically reliable sources, linguistic and cultural barriers, and the limits of authenticity in digital replication—at a time when artificial intelligence carries both great promise and deep ambiguity.

2. CHATBOTS IN EDUCATIONAL AND CULTURAL SECTOR

The term ‘chatterbot’ was coined by Michael L. Mauldin in 1994 [1], though chatbot development began earlier. These computer programs are designed to simulate human-like conversations and respond intelligently based on predefined data structures [2]. Historically, early chatbot models such as ELIZA and ALICE demonstrated fundamental natural language processing capabilities, using pattern matching and rule-based responses. Contemporary AI-driven chatbots, including OpenAI’s ChatGPT and Google’s Bard/Gemini, leverage vast datasets and advanced machine learning techniques to generate contextually relevant responses [3]. A key distinction between the two is their training data—Bard/Gemini is trained on real-time internet-based text, providing more up-to-date information, whereas ChatGPT relies on curated dataset of books and articles, often leading to greater accuracy in factual responses [4].

In educational contexts, chatbots have been integrated into various learning environments, assisting students with homework, providing personalized learning experiences, and delivering immediate feedback [5] [3]. Additional advantages of AI chatbots for students include flexible personalized learning and skill development [3].

However, researchers caution that excessive over-reliance on AI-generated information may undermine students’ critical thinking and problem-solving abilities [5]. From an educator’s perspective, AI chatbots offer time-saving assistance and enhanced pedagogy, but human expertise remains essential for making informed pedagogical decisions. Teachers’ primary concerns include reliability, accuracy, fair assessment—particularly in written assignments—and ethical considerations [3]. To ensure the safe and constructive use of AI chatbots, institutions must adapt their policies and practices accordingly [6].

The integration of chatbot-powered educational tools in museums and cultural heritage institutions, though not yet widely adopted, is gaining momentum, particularly as a means of digital innovation, enhancing audience engagement, and improving museum management. These AI tools create a welcoming environment where visitors feel comfortable asking questions, making museum visits more engaging. They help build trust and encourage interaction, especially among young visitors and families.

Chatbots make museum visits more engaging by sparking curiosity and inviting visitors—especially those who are not experts—to explore exhibits more deeply. Through personalized and interactive conversations, they create a more immersive experience while also helping museums learn what truly interests their audience. Museums primarily use AI to reinterpret collections, encourage participatory storytelling, and support digital learning [7]. A key example is the European H2020 ReInHerit project, which developed an AI-based multimedia chatbot to improve visitor interaction and answer visitor questions about artwork content and context.¹

A group of authors introduced a context-aware chatbot system to overcome limitations in existing AI models, particularly their inability to provide accurate historical and cultural insights [8]. The authors developed the CHVQA (Cultural Heritage Visual Question Answering) dataset, which contains 2,890 cultural asset images and 55,000 question-answer pairs sourced from verified historical content. The open-source chatbot system offers museums a scientifically accurate tool to enrich visitor experiences. The authors also emphasize the need for clear regulatory compliance to ensure ethical AI usage in cultural heritage. Chatbots should rely on curated knowledge from museum experts rather than solely on AI-generated responses in order to minimize the risk of misinformation. Additionally, obtaining user consent before data collection and ensuring secure data storage are crucial for maintaining visitor trust.

¹ <https://reinherit-hub.eu/tools/apps/>



In Serbia, the first and only chatbot currently in use exclusively in the higher education sector is ADA, developed by the Belgrade Business and Arts Academy of Applied Studies [9]. Initially designed to assist with communication and e-learning administration during the COVID-19 pandemic, ADA'S but purpose was subsequently broadened to include facilitating exam registrations and providing instant academic support through integration with Viber and Facebook Messenger. To date, no chatbots have been implemented in the educational or cultural heritage sectors in primary and secondary schools, museums, galleries, or endowments across Serbia. The goal of this project was to develop chatbots with practical applications within these institutions. In classrooms, the use of the chatbot goes beyond administrative tasks, as it can actively assist both students and teachers in learning and teaching processes, making teacher-student interaction more engaging. Within cultural institutions, the chatbot can serve as virtual guides, providing historical context, answering questions, and personalizing the visitor experience. It can also offer information in multiple languages, making cultural heritage more accessible to international visitors. Another potential application is its integration into special programs, where it can be used in interactive storytelling or simulations of historical events.

2.1. DIGITAL DOPPELGÄNGERS

Digital doppelgängers of individuals can be categorized into three types [10]: (1) historical digital doppelgängers – chatbots that enable communication with significant figures from the past (e.g., DigiDan), (2) griefbots – chatbots designed to simulate deceased individuals for the purpose of remembrance and emotional support (e.g., HereAfter AI), and (3) special-purpose digital doppelgängers—chatbots that play a crucial role in cultural memory, such as preserving the voices of Holocaust survivors, ensuring their testimonies remain accessible even after they are gone. Since DigiDan is the closest existing chatbot to the ones we propose, this paper will not focus on the second and third category of digital doppelgängers but will instead draw a parallel with it.

In 2023, DigiDan, a chatbot trained on the works of philosopher Daniel Dennett with his approval, was created. Using a fine-tuned GPT-3 model, DigiDan replicated Dennett's writing style, reasoning, and humor. A study with over 400 participants found that even professional philosophers struggled to distinguish the

chatbot's responses from Dennett's actual words [11]. Notably, Dennett was still alive when the chatbot was developed but passed away shortly after its release, raising important questions about digital immortality and the role of AI in preserving intellectual legacies.

Within the first category of doppelgängers, which includes DigiDan, we propose a subgroup of cultural-educational digital doppelgängers—chatbots designed for educational purposes and the promotion of cultural heritage within cultural institutions. Our pilot projects involve the Nikola Tesla and Branislav Nusic chatbots—virtual representations of Serbia's distinguished historical figures: Tesla as a renowned scientist, engineer, futurist, and inventor, and Nušić as a prominent playwright, satirist, essayist, novelist, and founder of modern rhetoric in Serbia. Developed for long-term integration into both the educational system and cultural sector, these chatbots aim to make historical personalities more accessible and engaging by authentically articulating their ideas, lives, and creative contributions. Additionally, the chatbots have been developed for inclusion in a broader platform, envisioned as a central digital space where various Serbian historical figures will be interactively available, facilitating enriched learning experiences and cultural exploration.

Two research studies are currently being conducted to assess the effectiveness and application of both Tesla and Nušić chatbots. In classrooms, these chatbots aim to help students learn about history, literature, and science in an engaging, interactive way. The Branislav Nušić chatbot can primarily enhance Serbian language and literature classes by offering analyses and explanations that deepen understanding of his literary works, especially his comedies and satires. Meanwhile, the Nikola Tesla chatbot provides detailed information about his inventions, patents, and scientific contributions, making it suitable for integration into subjects such as Physics, History, Serbian language and Literature, and English language. Currently available in both Serbian and English, the Tesla chatbot is also valuable in bilingual educational settings.

In cultural institutions, interactive chatbot communication enriches visitor experiences. The Nikola Tesla chatbot is primarily intended for the Nikola Tesla Museum in Belgrade but can also be used in other institutions such as the Museum of Science and Technology or various Tesla endowments nationwide. Similarly, The Branislav Nušić chatbot would serve institutions like the Branislav Nušić Endowment, the Branislav Nušić National Library, or the Branislav Nušić Theatre.



Importantly, their use is not limited to institutions explicitly bearing their names; the chatbots could effectively enhance other cultural and educational spaces aligned with their respective themes. This approach would promote Serbian cultural heritage and scientific achievements globally, by presenting them in an engaging and modern format. Through interactions with local and international visitors, these chatbots can significantly enhance the museum experience and foster deeper appreciation for Serbia's national heritage.

3. CREATING NIKOLA TESLA AND BRANISLAV NUŠIĆ CHATBOTS

3.1. TESLA'S CHATBOT: CHALLENGES IN REPLICATING A SCIENTIFIC GENIUS

Developing a chatbot to represent Nikola Tesla proved significantly easier compared to Branislav Nušić, primarily due to the abundance of information available about Tesla's life, work, and inventions. Biographies, scientific articles, patents, and letters related to Tesla are available in multiple languages and from diverse sources, providing a richer corpus for training the chatbot. However, an extensive dataset does not necessarily guarantee authenticity. Numerous myths and legends surrounding Tesla prompt critical questions: Which sources are reliable? How can genuine information be differentiated from sensationalist stories? These issues necessitate meticulous selection and validation of data.

Another significant challenge is the language or dialect Tesla used. Tesla spoke in the Ijekavian variant of the Serbian language, requiring special linguistic adaptations within the chatbot. This poses particular difficulties when adapting content accurately from the perspective of speakers predominantly familiar with the Ekavian variant. Furthermore, Tesla employed terminology typical of his era, which required additional effort to adapt accurately. Linguistic assistance and meaningful comprehension of historical context are essential for creating an authentic digital representation of Tesla.

A more complex issue is Tesla's multilingual abilities. Historical accounts indicate he spoke nine languages, yet details on his proficiency in each are limited. If the chatbot were ever to converse or write in multiple languages, it would be necessary to determine whether Tesla was truly fluent or only superficially familiar with each language. Ironically, AI naturally improves with training, meaning the chatbot might eventually surpass Tesla's own command of languages, such as German.

This scenario raises fundamental questions: Would the chatbot remain authentic representation if it surpasses Tesla's linguistic capabilities? Furthermore, if the chatbot is to communicate realistically, more details on Tesla's speech patterns and possible accents when speaking foreign languages would be essential.

Another challenge is the historical authenticity of Tesla's language and terminology. Replicating Tesla's voice requires the use of the historically accurate linguistic expressions and stylistic nuances of his time. Since historical records do not provide exhaustive insight into Tesla's proficiency in each of the nine languages he spoke, careful consideration is needed when determining the chatbot's linguistic capabilities.

The irony deepens: due to machine learning, the chatbot could eventually surpass Tesla's original linguistic skills. Would such a chatbot remain truly authentic? Is a digital doppelgänger a faithful representation if it acquires knowledge the original person never possessed? These questions raise profound philosophical and ethical dilemmas about authenticity and identity in the digital age. If an AI-driven chatbot develops skills and knowledge beyond what the historical figure originally had, can it still be considered a faithful replica—or is it instead an evolved version of the original personality? Does authenticity require adherence strictly to historical fact, or can it be grounded in an idealized potential of the reconstructed individual?

The problem becomes even more complex when considering subjective interpretations: if Tesla himself were alive today, would he recognize this chatbot as his digital heir, or see it as an interpretation shaped by modern technology and others' expectations? Ultimately, the critical question remains whether a digital doppelgänger should remain a static reconstruction of a particular historical era, or evolve into something the original figure could never have become—but which, under different circumstances, might align with their vision and potential.

3.2. NUŠIĆ'S CHATBOT: CHALLENGES IN DIGITIZATION AND CONVEYING HUMOR

Creating a chatbot inspired by Branislav Nušić presents a unique set of challenges. Although Nušić's literary works left an enduring mark on Serbian culture, they are neither widely digitized nor extensively translated. This gap poses a significant obstacle in building a comprehensive dataset needed for chatbot training.



To overcome this challenge, efforts have begun to digitize Nušić's writings, including analyses and critiques by other authors. While essential for the chatbot's authenticity, this effort also highlights a broader societal issue: the inadequate digitization of cultural heritage. Ideally, national libraries and cultural institutions should undertake this task, yet the responsibility has largely fallen to individuals attempting to revive cultural heritage through modern technology.

In contrast to the extensive data available for Tesla, the case of Nušić is starkly different. Many of his works are not yet digitally available, limiting material crucial for chatbot training. Nušić's unique style and humor are intimately tied to a specific historical period. If the chatbot utilizes contemporary language, can it still be considered faithful to the original? The lack of digital resources points to the broader issue of cultural heritage digitization.

Another critical challenge is accurately conveying Nušić's humor. Humor is strongly influenced by cultural and historical context. If the chatbot employs contemporary language, it risks straying from authentic stylistic expression. Language is not merely a communication tool; it reflects the spirit of the era—sentence rhythms, vocabulary choices, and subtle meanings that shape humor. Nušić's style was rich in archaic expressions, bureaucratic language satire, and specific social contexts that informed his comedy. If the chatbot instead uses modern expressions, it risks sounding like a 21st-century author imitating Nušić rather than genuinely replicating him.

Conveying context is another issue. Nušić's humor was not simply wordplay but a sharp critique of society, satirizing particular political and social conditions—bureaucracy, provincial mentality, societal hypocrisy. While a chatbot could recognize patterns in Nušić's writing, it would struggle to fully comprehend the deeper motivations behind his jokes and their historical relevance. Replacing these satirical moments with generalized humor would dilute their essence and strip them of nuance.

Nevertheless, could such a chatbot still be useful? If seen as a tool to adapt Nušić's humor for contemporary audiences, it could generate a new interpretation of his style, adjusted to modern realities. Rather than literal reconstruction, the chatbot could employ Nušić's method of identifying and satirizing modern-day absurdities, thus preserving the spirit of his satire. The question remains: Would this chatbot be a true digital representation of Nušić, or rather something entirely new, inspired by his legacy yet inevitably different—a digital doppelgänger with a life and purpose of its own?

4. ETHICAL DILEMMAS

Throughout the development and testing of our pilot projects, several ethical dilemmas have surfaced, particularly concerning authenticity, knowledge boundaries, and ethical use of AI. Creating digital replicas such as those modeled after Nikola Tesla nad Branislav Nušić raised important questions about uniqueness of human experience. If historical figures can be digitally reconstructed, does it diminish their authentic human essence [12]? While our chatbots are constantly trained to mimic linguistic styles and speech patterns as much as possible, they inherently lack consciousness, genuine emotional responses, and authentic thought processes. For instance, can a chatbot truly replicate Nušić's humor, or is it limited to recombining existing comedic content? Moreover, despite carefully replicating speech, our chatbots are fundamentally unable to authentically engage with personal or emotionally complex questions.

Another issue arising from our projects concerns the authenticity of historical representation when contemporary data is incorporated. Are our digital replicas historically accurate, or do they represent modernized versions? Ethical challenges become particularly evident when programming the chatbots to simulate emotions. For instance, is it ethical for the Tesla or Nušić chatbot to mimic expressions of sadness or joy without genuine emotional experiences? Furthermore, we confronted the sensitive dilemma of how chatbots should interact with descendants of these historical figures versus general visitors or admirers.

Authenticity of knowledge presented another significant challenge. Historical figures like Tesla or Nušić had clear knowledge limitations. For example, questions arose during testing such as whether Tesla knew who the king of Norway was in 1940—a fact easily retrieved today, but perhaps never known by Tesla himself. Providing answers beyond the figure's historical knowledge base compromises authenticity and credibility. Moreover, how should chatbots handle historical uncertainties or unknowns? Unlike humans, who naturally forget or selectively recall information, AI has a comprehensive, unfiltered access to data, complicating the distinction between confirmed historical knowledge and speculative information.

AI's unpredictable behavior has been another ethical concern in our pilot tests. Often described as a "black box" phenomenon, AI's decision-making processes can be opaque, making it challenging to understand why a chatbot generated a particular response. Although this



sometimes leads to novel insights, it equally introduces potential errors, biases, and incorrect information, particularly problematic in educational contexts [13]. AI inference relies on statistical probabilities rather than genuine understanding, fundamentally differentiating chatbot reasoning from human cognition.

Bias also posed a practical ethical dilemma in our projects. Despite careful consideration, creators and trainers inevitably influence chatbot responses, potentially embedding their own biases into digital replicas of historical figures. Additionally, privacy emerged as a key issue. During testing, we noted users' uncertainty regarding how their interactions with chatbots might be recorded, stored, or accessed by third parties. These concerns were compounded by the lack of comprehensive data regulation and risks associated with the misuse of sensitive information [14] [15].

Reliability posed further challenges. The historical sources—letters, memoirs, newspapers, and literary works—are inherently subjective and limited, complicating efforts to ensure historical accuracy in our chatbots. For the Tesla and Nušić chatbots, publicly available materials were supplemented by consultations with Tesla's biographer, Milovan Matić, and historian Lieutenant Uroš Jovanović, who provided valuable assistance regarding the history of the Serbian Military Frontier (Srpska vojna Krajina).

Legally, our use of public-domain materials, such as the works of Branislav Nušić, is permissible, but significant ethical concerns still remain. For instance, we need to thoughtfully consider how the descendants of Tesla and Nušić might perceive the creation of digital replicas of their ancestors being created. It is essential to recognize and respect their perspectives, especially concerning consent, privacy, preservation of family legacy, and our ethical responsibility toward these families. Thus, even when the use of creative or historical content is legally acceptable, it can still raise serious ethical questions if explicit consent is not obtained. This legal and ethical grey area clearly requires further discussion and calls for establishing clear regulatory guidelines.

5. CONCLUSION

A critical limitation identified in our pilot projects is the inability of educational chatbots to fully replicate the emotional support and mentorship provided by human educators [3]. To maximize AI's educational potential while mitigating its shortcomings, we recommend that institutions adopt preventive measures informed by our experience—raising student awareness about digital inequality, AI reliability, and ethical implications, alongside offering professional development opportunities for educators. Training should focus on integrating AI as a complementary tool rather than a replacement for human instruction, thereby nurturing students' critical thinking and problem-solving skills.

Our pilot experiences also demonstrated the importance of ongoing chatbot refinement in cultural context. Incorporating visitor feedback significantly improves chatbot accuracy, engagement, and inclusivity. Additionally, ethical challenges, such as transparency, data privacy, and misinformation [8], must be proactively managed. Our projects have stressed the need for balancing technological innovation with cultural authenticity, ensuring chatbots enrich rather than distort historical narratives.

Creating digital doppelgängers like the Nikola Tesla and Branislav Nušić chatbots extends beyond technical innovation; it pushes us to reconsider our understanding of historical authenticity, cultural interpretation, ethical responsibility, and identity. As virtual replicas become more sophisticated, it remains crucial, as our pilot projects illustrate, to view these technologies as tools for enhancing educational and cultural experiences—not as substitutes for genuine human legacies. Rather than striving for perfect replicas, our aim should be developing innovative tools that help us better appreciate and understand ideas and contributions of the great minds.



REFERENCES

- [1] G. Molnár i Z. Szűts, „The Role of Chatbots in Formal Education. *IEEE 16th International Symposium on Intelligent Systems and Informatics – SISY 2018* (2018): 197-202. <https://doi.org/10.1109/SISY.2018.8524609>,” u *IEEE 16th International Symposium on Intelligent Systems and Informatics – SISY 2018* (2018).
- [2] A. Khanna, B. Pandey, K. Vashishta, K. Kalia, B. Pradeepkumar i T. Das, „A study of today's AI through chatbots and rediscovery of machine intelligence,” *International Journal of u-and e-Service, Science and Technology*, t. 8, br. 7, pp. 277-284, 2015.
- [3] L. Labadze, M. Grigolia i L. Machaidze, „Role of AI chatbots in education: Systematic literature review,” *Int J Educ Technol High Educ*, 2023.
- [4] J. Rudolph i S. Tan, „War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education,” *Journal of Applied Learning and Teaching*, t. 6, br. 1, 2023.
- [5] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann i E. Hüllermeier, „ChatGPT for good? On opportunities and challenges of large language models for education,” *Learning and Individual Differences*, t. 103, p. 102274, 2023.
- [6] D. Baidoo-Anu i L. Owusu Ansah, „Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning,” *SSRN*, 2023.
- [7] The Directorate General for Communications Networks, Content and Technology (European Commission), „Opportunities and challenges of artificial intelligence technologies for the cultural and creative sectors,” Publication Office of the European Union, 2022.
- [8] P. K. Rachabatuni, F. Principi, P. Mazzanti i M. Bertini, „Context-aware chatbot using MLLMs for Cultural Heritage. (pp. 459-463),” u *Proceedings of the 15th ACM Multimedia Systems Conference*, 2024.
- [9] A. Vukomanović, N. Deretić, M. Kabiljo i R. Matić, „An example of chatbot in the field of education in the Republic of Serbia,” *Journal of process management and new technologies*, t. 10, br. 1-2, pp. 125-139, 2022.
- [10] Subotić, V., „O digitalnim blizancima i digitalnoj besmrtnosti,” <https://elementarium.cpn.rs/esaji/o-digitalnim-blizancima-i-digitalnoj-besmrtnosti/?script=lat>, 2025.
- [11] E. Schwitzgebel, D. Schwitzgebel i A. Strasser, „Creating a large language model of a philosopher,” *Mind & Language*, t. 39, br. 2, pp. 237-259, 2024.
- [12] J. Danaher i S. Nyholm, „The ethics of personalised digital duplicates: a minimally viable permissibility principle,” *AI and Ethics*, pp. 1-16, 2024.
- [13] J. Hutson, J. Ratican i C. Biri, „Essence as Algorithm: Public P Essence as Algorithm: Public Perceptions of AI-P ceptions of AI-Powered Avatars of Real People,” *DS Journal of Artificial Intelligence and Robotics*, t. 1, br. 2, pp. 1-14., 2023.
- [14] J. S. A. Karpus, „Persons and their Digital Replicas,” *Philosophy & Technology*, t. 38, br. 25, pp. 1-23, 2025.
- [15] C. Voinea, B. D. Earp, S. Porsdam Mann i J. Savulescu, „Digital doppelgängers, human relationships, and practical identity,” Pre-print: https://www.researchgate.net/publication/384254177_Digital_Doppelgangers_Human_Relationships_and_Practical_Id, 2024.



AN ANALYSIS OF THE USE OF ORTHOGRAPHIC RULES IN STUDENTS' WRITTEN ASSIGNMENTS IN DIGITAL ENVIRONMENTS

Saša Čorboloković*,
[0009-0002-3767-2313]

Valentina Gavranović
[0000-0002-3011-8769]

Singidunum University,
Belgrade, Serbia

Abstract:

Being literate in the 21st century implies more than knowing how to read, write, and calculate; in the digital era, literacy involves various aspects, including digital literacy. Information and communication technologies and the ever-growing amount and speed of written communication greatly affect how we use language in digital environments. This paper investigates orthographic rules and focuses on those segments specific to typing – namely, the use of a space, and writing the symbols that represent letters and punctuation marks. The main aim of this study is to identify and analyze common and recurrent mistakes undergraduate students majoring in English philology made in their essays written in English and translations written in Serbian. The results obtained from the analysis draw attention to those rules often violated in typing, and a possible language transfer in translation, focusing on contrastive analysis tenets. This study also draws attention to the significance of cultivating literacy in digital environments through the development of a neat and careful manner of using language in all contexts, particularly in formal and academic writing.

Keywords:

Orthographic Rules, Digital Literacy, Punctuation, Spacing, Language Transfer.

INTRODUCTION

Literacy has been defined and determined differently over time [1], [2]. Unlike the 20th-century interpretation of literacy as elementary, primary, or basic, incorporating reading, writing, and calculation, the 21st century introduces new aspects of the meaning of the term 'literacy'. Nowadays, literacy is defined more comprehensively, encompassing a wider and more complex scope, even changing the singular into plural word form; hence, we use the term literacies to identify different types, such as linguistic, mathematical, scientific, digital, media, environmental, financial, etc. [3]

The modern approach to literacy emphasizes the functional value of knowledge (called 'functional literacy') gained for solving various problems in everyday life [4]. With the development, expansion, and increasingly frequent use of modern information technologies in the modern world, digital literacy is not only becoming one of the indispensable

Correspondence:

Saša Čorboloković

e-mail:

scorbolokovic@singidunum.ac.rs



segments of literacy but is also slowly gaining equal status with elementary literacy. Digital literacy can be defined as a combination of knowledge, skills, and attitudes required for the efficient use of digital technology in all areas of contemporary life – education, business, responsible and active social participation, and free everyday, leisure activities. Furthermore, digital literacy also includes effective problem-solving processes, communication, information management, collaboration, content creation, and sharing; it also implies knowledge construction [5]. The importance of developing students' digital literacy at the tertiary level has been emphasized; researchers investigated the issue from various perspectives: building a professional digital identity [6], preparing students for the 21st-century real-life demands [7], and developing students' transversal skills and competencies necessary for the 21st-century workplace [8].

In addition to the aspects mentioned earlier in this paper, digital environments encompass written communication, which requires correct and efficient language use in digital format. This paper aims to investigate the commonest mistakes students make in writing in a digital format, particularly focusing on the use of punctuation and spacing in typing. As regards punctuation, this paper primarily investigates those aspects differently realized in Serbian and English, relying on the main principles applied in contrastive linguistics. This study draws attention to the development of educated literate people who efficiently and correctly use language in pervasive digital environments incorporating formal and informal registers.

2. ORTHOGRAPHY IN THE DIGITAL AGE

Orthography is an indispensable part of literacy. Nonetheless, the pace and manner of modern communication in the digital age highly affects proper and correct writing. The application of orthographic rules in a digital environment raises several questions and dilemmas: do they belong to a special type of orthography, or do they represent a standard form used in the digital environment; what are the basic rules and what methodology should we employ in teaching; what subject/subjects should contain lessons focusing on orthography, and how it can be systematically presented and acquired.

The term 'digitalni pravopis' is defined in the Serbian language as a set of rules or guidelines defining the correct use of language in digital communication (e-mail, writing on the Internet, social networks, forums,

blogs, and various messaging applications in general). This area is strictly defined by rules stated in several books and guidelines defining orthographic conventions in the Serbian language. They rely on numerous rules prescribed for writing in general and additional rules that apply to printing. Although they have existed since the invention of the printing press, these rules nowadays gain more prominence because of the pervasive writing needs and habits in digital environments, particularly enhanced by the use of social media. On the other hand, there is no single equivalent phrase in English, which indicates the absence of strictly defined rules that are specific only to the digital environment. Generally speaking, orthographic rules used in digital environments encompass the application of all the orthographic rules used in handwriting, with an addition of several straightforward rules: the correct typing of spaces and punctuation marks (periods, commas, question marks, exclamation marks, colons, brackets, dashes, hyphens, etc.). In other words, when writing in digital environments, it is essential to respect typographic rules, integrating traditional grammar and orthographic conventions from Serbian or English, alongside specific typographic guidelines. [9].

The complexity of communication types in the digital age – varying from formal, professional, and academic registers to rather informal contexts – influences the way we apply orthographic rules in typing. Namely, the informal register (everyday colloquial language used in personal contexts) can contain a certain degree of flexibility and creativity, often featuring abbreviations and various emoticons. Although abbreviating words, phrases, and sentences follow certain rules, and one's arbitrary abbreviating is not allowed, nor recommended, sometimes these rules are influenced by the limitations of some platforms (on the H/Twitter network there are abbreviations, and hashtags are common due to the limited number of characters). The degree of application of orthographic rules in a digital environment also depends on one's attitude and perceptions of the platform they use. For example, common and standard orthographic rules are expected to be applied on LinkedIn because it represents a professional context and individuals having profiles there strive to present themselves well and respectfully, highlighting the main features of their professional biography and professional attitude (which can be enhanced by proper use of language). On the other hand, the main approach to the use of language in social media platforms, such as Instagram and Facebook, can greatly differ because it is perceived as informal, light-hearted, and belonging to leisure time, reserved for a



break from professional constraints. Formal communication, including a wide range of contexts – such as professional, and academic registers – demands the use of language that indicates the level of education, expertise, professional and personal attitude toward the intended audience and the topic addressed in the written communication. Unlike printed versions of books and academic papers, which usually go through a strict proofreading and editing stages, many written texts found online do not have any monitoring process which results in the approval for publishing. Therefore, we can find numerous examples of incorrect uses of language in written evidence online, which serves as a source of further incorrect instances of language. In such a way, mistakes are repetitively made, copied, enhanced, and after negligence and common use of such blunders become widely accepted and, therefore, perceived as norms. Hence, it is important to focus on this aspect of writing, by introducing a systematic way of teaching and using proper orthographic conventions in the digital, both formal and informal environments.

Orthographic rules used in digital environments are adopted at different stages throughout education and further professional development. Students first encounter these rules mainly at school, during their formal education; therefore, the school should be the first and foremost foundation for learning and applying these rules, thus forming the habits of correct typing. The greatest responsibility lies with education policies focusing on a systematic approach to teaching this aspect of writing and literacy, involving teachers, educators, and instructors as the mediators for their implementation. As regards the Serbian education system in primary and secondary schools, lessons devoted to proper typing and orthographic rules in a digital environment can be found only in the latest editions of the Serbian language and Information, Communication, and Technology textbooks for primary and secondary levels. The practice of writing assignments as a part of evaluation, particularly in Serbia and foreign language classes, is mainly conducted through handwriting, and students usually do not practice the use of orthographic rules when writing in digital environment; in other words, students do not have much opportunity to type formal assignments that are assessed during their formal education.

Additionally, courses (such as Modern Business Correspondence, Office Management, etc.) taught in some vocational schools of economics and law in Serbia have contained lessons devoted to proper typing for decades – still, with no specific mention of the term 'digitalni pravopis'. Furthermore, in certain professions, such

as journalism, marketing, PR, and the IT industry, the proper application of orthographic rules in digital format is a requirement, often implied, representing a prerequisite for successful and professional communication with consumers, colleagues, and partners.

However, the reality is different, and, due to the prevalence of digital language input and the absence of critical thinking, errors made in writing are numerous, contributing to the adoption and creation of a careless and sloppy style characterized by erroneous and illiterate text. Unfortunately, the consequences of general carelessness and insufficient awareness of the importance of proper application and use of orthographic rules in typing are also present in professional contexts, even at the academic level. Therefore, it is of uttermost importance to understand the relevance of correct language usage and the application of orthographic rules and to instruct young people for effective, clear, and linguistically literate communication in a rapidly changing world.

3. RESEARCH METHODOLOGY

This paper analyzes how well undergraduate students majoring in English philology apply orthographic rules in writing in digital environments and investigates mistakes made in terms of punctuation and spacing. The main aim of this study is to focus on the commonest mistakes and draw attention to a possible language transfer and interlingual influence.

The analyzed corpus comprises 126 student papers – 67 essays written in English and 59 translation tasks written in Serbian. Students were not instructed to apply any specific orthographic and typing rules; they focused on the essay structure, topic development, coherence and language precision, and the most appropriate translation equivalents in essays and translation tasks, respectively. The authors aimed to analyze those aspects of writing usually taken for granted and, therefore, did not explicitly instruct the students to focus on them, thus eliciting the desired results.

The method used in this study is quantitative and relies on error analysis. Common mistakes were identified and classified into two categories – punctuation and spacing in Serbian and English.

Regarding punctuation, the authors focused only on several aspects: the use of punctuation marks in direct speech, writing diacritics, and punctuation marks. This study aims to investigate only those aspects of orthographic conventions that are specific to typing, while general rules were not the focus of this analysis. In other



words, the analysis of general orthographic rules were not the focus in this study because the authors aimed to excerpt only those specific examples related to typing, and, thus, draw attention to a neat and careful use in everyday communication and, particularly, in formal contexts, encompassing written assignments at the university level, done by future professionals in language use and analysis.

4. RESEARCH RESULTS AND DISCUSSION

The essays written in English and translation tasks were first analyzed and the use of orthographic conventions were categorized according to the incorrect applications noticed in students' written assignments. The mistakes identified in students' papers can be classified into three areas: the use of a space, the use of symbols, and the use of punctuation.

Firstly, the analysis shows some common recursive mistakes found in students' writing in both languages, and they relate to the use of a space (a blank area used in print, separating words, clauses, and sentences). In this analysis, we combined both languages because there are no differences between the use of a space in Serbian or English. Namely, students consistently used the same way of making a space in both languages, and there are no differences identified in essays or translation tasks.

In general, the analysis shows the following common mistakes in leaving a space in Serbian and English: spaces between words and punctuation marks; the use of space within parenthesis; the use of a space before and after a dash.

1. Spaces between words and punctuation marks. When typing, after punctuation marks (a comma, question, and an exclamation mark, colon, semicolon, ellipsis) there should be one space. Mistakes identified in students' papers relate to the absence of a space after these punctuation marks (the example that illustrates such a mistake is as follows: *Despite being so naive,he...*), instead of using a space after the comma and before the word 'he': *Despite being so naive, he...*). On the other hand, what was identified as a repetitive mistake relates to leaving a space before a punctuation mark (especially before quotation marks at the beginning of the quote, usually accompanied by a space before the quotation mark at the end of the quote; an example of the incorrect usage is as follows: " *She finally decided to leave, but...* "). Another mistake identified in one paper

written in English relates to the use of a space after a word and before an ellipsis (the example is as follows: *Then I realized how much time, energy and money I have wasted, but ...*).

2. The use of a space within parenthesis. When typing, there is no space before the first and last word within parenthesis. Leaving space in such situations was also identified in a high percentage of translation tasks (almost 45%) that contained such examples in the original text in English. However, it is not certain whether those students who did not make a mistake knew the rule or just copied the way it was written in English when translating the text from the parenthesis into Serbian (the example that illustrates such a mistake is as follows: *Across the UK, there are five stages of education: early years, primary, secondary, Further Education (FK) and Higher Education (HE)*, and the translation that illustrates the incorrect usage of a space within parenthesis is: *U Velikoj Britaniji postoji pet nivoa obrazovanja: ranom osnovnom srednje, više(FK) i visoko obrazovanje (HE)*).
3. The use of space before and after a dash. In typing, there is a space before and after a dash. However, the analysis shows that students occasionally omitted a space and merged words with a dash (the example illustrating such a mistake is as follows: *The relevance of such theory-even not finally proven-is evident not only in science but everyday as well*).

Furthermore, the analysis also reveals inconsistency in the use of spacing in typing; namely, the majority of students used spacing differently throughout the same essay or translation task. Only 12 papers showed that these mistakes were recurrent, which brings to the conclusion that these students, with a high degree of certainty, did not know the rules because they consistently applied the opposite in their writing.

Secondly, the analyzed written assignments contained many mistakes related to the use of symbols, in both languages. The analysis of these mistakes shows two tendencies: the incorrect use of letters in the Serbian language, and the incorrect use of symbols representing punctuation marks (these relate to the use of a dash (–), and quotation marks. The latter relates only to Serbian because the quotation marks were found only in translation tasks; there was no direct speech used in the essays written in English, and, therefore, no reliable conclusion can be made on the use of quotation marks in English.



The following examples illustrate the commonest mistakes made in students essays written in English and translation tasks written in Serbian as regards the use of symbols.

1. These mistakes primarily relate to the way letters in Serbian were typed. 55 translation tasks were typed in Latin alphabet and only 4 in Cyrillic. Those typed in Cyrillic consistently used proper letters. However, common mistakes were identified in translation tasks typed in Latin; namely, these include the absence of diacritic symbols, and students typing wrong or non-existent letters, such as 'c' for 'ć' or 'č', 's' or 'sh' for 'š', 'dj' for 'đ', and 'dz' for 'dž'. Therefore, common mistakes found in students translations are as follows: 'ceta' instead of 'četa', 'skola' instead of 'škola', 'cup' instead of 'ćup', 'sharm' instead of 'šarm', 'dzangrizav' instead of 'džangrizav', and many other examples illustrating the incorrect use of letters.
2. In the majority of papers, in both languages, students did not use the symbol for a dash (–), but a hyphen (-); only in 7 papers, the symbol for a dash was consistently used. In translation tasks, students already had the dash used in the original, English version, used correctly as a symbol (–), but only in 7 papers the symbol was correct; other papers showed the use of a hyphen (-) instead. In essays written in English, 5 students used a dash as a punctuation mark, but in all these papers the symbol was mistyped (a hyphen used instead of a dash). The example illustrating this common mistake in students translation tasks is as follows: *People who possess intrapersonal intelligence are generally aware of themselves, their strengths and weaknesses – they are associated with introspection and self-reflection*, and the translation that illustrates the incorrect usage of a space within parenthesis is: *Osobe koje imaju intrapersonalnu inteligenciju uglavnom su svesne sebe, svojih vrlina i mana - povezane su sa introspekcijom i autorefleksijom*.

Finally, the last aspect analyzed in students' papers relates to the use of punctuation marks in direct speech in the Serbian language in translation. The commonest mistake the students made involved the use of quotation marks – instead of typing „”, in almost 70% of papers, students used English symbols " ". Additionally, students often omitted a comma after the quotation mark (e.g., instead of typing „Zar niste rekli da nemate

ništa sa tim?”, *pitala me je naposletku*, many students wrote "Zar niste rekli da nemate ništa sa tim?" *pitala me je naposletku*). The identified mistakes comply with the previous research results that relate to the use of punctuation marks in reported speech [10].

The analysis of students' written assignments – essays written in English and translations written in Serbian – shows several tendencies regarding the proper use of orthographic rules in typing. The first relates to the common mistakes made in both languages regarding the use of a space in typing. Namely, recurring mistakes included the incorrect connection of words and punctuation marks, such as when words followed punctuation marks (a comma, full stop, question mark, or semi-colon). On the other hand, students wrongly separated quotation marks from words within. Another repetitive mistake identified in students' translations from English into Serbian relates to the symbols used in typing – the absence of diacritic symbols in Serbian Latin, which indicates that the students used the English keyboard instead of Serbian. Apart from these symbols, other common mistakes found in students' translations include quotation marks as used in the English language – the tendency illustrates the influence of the English language on Serbian. Finally, the repetitive mistake relates to the use of a hyphen instead of a dash, which also illustrates the use of keyboard that does distinguish between the two symbols, leaving space for a mistake. Even though the outlook of a dash and a hyphen in Serbian is clearly defined and explained, there is no such consistency in the English language and it is not rare to find these two symbols interchangeably used even in formal and academic contexts, including language and linguistic-related papers.

5. CONCLUSION

This study focused on one aspect of literacy in digital environments – understanding and properly applying those orthographic rules specific for typing – spacing and the use of punctuation and symbols in writing. The aim of the research was to identify the commonest mistakes students made in typing when writing essays and translation tasks in the Serbian and English languages, respectively.

Although the need for efficient information exchange may justify some instances of incorrect application of orthographic conventions, it is crucial to emphasize the importance of cultivating proper language use in digital environments – such an approach contributes to the



development of literacy perceived in its broadest sense. David Crystal explains this phenomenon as a result of language economy and speed of communication, calling this phenomenon the principle of "save a keystroke" [11]. This study draws attention to the following aspects that should be considered when typing: respecting all orthographic rules defined for all writing environments (in handwriting and typing), using spacing properly, and using the Serbian Latin keyboard to mark letters such as 'ć', 'č', 'š', 'đ', 'ž', and 'dž'. As Piper and Klajn pointed out, an educated person knows all capital and small letters in a language – actively, to read them correctly, and passively, to write or type them [12].

Therefore, in order to develop students' literacy for the 21st century, it is important to teach students not only how to write but also how to type properly, respecting all prescribed orthographic rules. The way we write in a digital environment often says a lot about us, our level of education, literacy, and the approach we have toward work, profession, and leisure time. Early and systematic adoption of careful and correct writing will create a habit of the educated writing style; on the contrary, carelessness leads to sloppiness and an illiterate manner of communication.

REFERENCES

- [1] H. Popović, *Pismenost pod globalnom lupom*, Beograd: Institut za pedagogiju i andragogiju Filozofskog fakulteta u Beogradu i Društvo za obrazovanje odraslih, 2019.
- [2] S. Antić and J. Stevanović, „Kako unaprediti razumevanje pročitano: mapiranje kritičnih zona za pedagoške intervencije,“ *Nastava i vaspitanje*, t. 73, br. 3, p. 345–362, 2024.
- [3] B. Cambridge, „Learning, Knowing, and Reflecting: Literacies for the Literacies for the 21st Century,“ *International Journal for the Scholarship of Teaching and Learning*, t. 1, br. 2, p. 1–7, 2007.
- [4] S. Tomić, „Knjiga koja nedostaje: priručnik za funkcionalnu pismenost”, Beograd: 3D+, 2022.
- [5] R. Vuorikari; Y. Punie, S. Carretero Gomez and G. van den Brande, *DigComp 2.0: The Digital Competence Framework for Citizens. Update Phase 1: the Conceptual Reference Model*, Luxembourg: Publications Office of the European Union, 2016.
- [6] H. Četojević, „The Impact of Digital Literacy on the Development of Higher Education,“ *Pedagoška stvarnost*, t. 69, br. 2, p. 189–192, 2024.
- [7] A. I. Santos and S. Serpa, „The Importance of Promoting Digital Literacy in Higher Education,“ *RedFame Publishing*, t. 5, br. 6, p. 90–94, 2017.
- [8] N. Khan, A. Sarwar, T. B. Chen and S. Khan, „Connecting Digital Literacy in Higher Education to the 21st Century Workforce,“ *Knowledge, Management & E-Learning*, t. 14, br. 1, p. 46–61, 2022.
- [9] S. Čorboloković and V. Gavranović, *Pravopis u kontrastu*, Beograd: Univerzitet Singidunum, 2024.
- [10] S. Čorboloković and V. Gavranović, „Komparativno-kontrastivna analiza upotrebe interpunkcijskih (i pravopisnih) znakova u srpskom i engleskom jeziku,“ *Zbornik radova Filozofskog fakulteta u Prištini*, t. 54, br. 1, p. 67–84, 2024.
- [11] K. Crystal, *The Cambridge Encyclopedia of the English Language*, Cambridge: Cambridge University Press, 2022.
- [12] P. Piper, I. Klajn and R. Dragičević, *Normativna gramatika srpskoga jezika (četvrto izmenjeno izdanje)*, Novi Sad: Matica srpska, 2022.



DIGITAL PLATFORMS FOSTERING INTERCULTURAL COMPETENCE IN LANGUAGE CLASSES

Ivana Đerić*,
[0000-0003-2914-932X]

Neda Maenza,
[0000-0002-4974-6052]

Nataša Stanišić,
[0000-0001-9487-8225]

Aleksandra Gajić,
[0000-0002-6804-4662]

Tijana Gajić
[0000-0002-5013-2867]

Singidunum University,
Belgrade, Serbia

Abstract:

The increasing interconnectedness of the globalized world has underscored the necessity of fostering intercultural competence alongside language learning. In this context, the development of intercultural communication skills is central to promoting tolerance, understanding, and empathy among diverse cultural groups. Foreign language proficiency is crucial in overcoming linguistic and cultural barriers, yet the path to achieving intercultural competence remains complex and often secondary to traditional educational goals. This paper explores the role of digital platforms in enhancing both language proficiency and intercultural competence in foreign language classes at Singidunum University, with a particular focus on French, Spanish, and English courses. The integration of digital platforms into these language programs has proven to be an effective tool for bridging cultural divides. These platforms facilitate not only the development of linguistic skills but also increase cultural awareness by providing authentic, interactive, and cooperative learning environments. Digital platforms enable students to engage with people coming from diverse cultural backgrounds thus dismantling stereotypes, promoting cross-cultural understanding, and nurturing critical intercultural skills such as communication, critical thinking, and empathy. This overview paper demonstrates how technology can transform foreign language education into an engaging and culturally inclusive experience, enhancing the global competence of students.

Keywords:

Intercultural Competence, Language Learning Platform, Cross-Cultural Communication, Language Courses, Educational Technology.

INTRODUCTION

Language learning extends beyond acquiring linguistic skills. It connects individuals across cultures, fostering dialogue, mutual understanding, and empathy. It empowers young people to communicate across cultural boundaries, breaking down social and ethnic barriers while enhancing intercultural awareness.

Intercultural understanding is essential in dismantling stereotypes and building trust among communities. The exchange of ideas and traditions promotes inclusive dialogue, contributing to a more tolerant society. When combined with cultural appreciation, language learning has the power to transform individual's perspectives and strengthen global citizenship.

Correspondence:

Ivana Đerić

e-mail:

idjeric@singidunum.ac.rs



Rapid technological advancements have made it possible for information and communication technology to be integrated into education as a powerful catalyst for societal change. Digital platforms have revolutionised language learning by fostering intercultural competence through interactive experiences.

The paper explores the role of digital platforms in enhancing intercultural competence in French, Spanish, and English courses taught at Singidunum University. It highlights best practices for integrating digital tools into language education so as to prepare students for meaningful cross-cultural interactions by analyzing academic perspectives and real-world applications.

2. LITERATURE REVIEW

The increasing interconnectedness of the world has reinforced a common goal within the international community: the development of effective intercultural communication skills [1]. This evolving paradigm has significantly influenced foreign language education, shifting its focus toward the development of cultural competencies. It encourages an appreciation for inclusive values while fostering tolerance, respect, and openness toward different cultures. Furthermore, it facilitates the exchange of ideas and experiences across cultural boundaries. As Petrović [2] notes, effective communication with individuals from diverse cultural backgrounds is a skill that requires continuous development. A key aspect of this ability is intercultural sensitivity, which entails recognizing significant cultural differences and cultivating positive attitudes toward them.

The role of educators is crucial in developing intercultural competence among students, both within national contexts and on a global scale. Teachers must not only possess these skills themselves but also be capable of effectively transferring them to their students. Despite the implicit presence of cultural elements in education, intercultural competence remains a secondary objective rather than a core educational priority. There is no universally established approach for its integration, making its incorporation into formal education a gradual and complex process [3].

Foreign language proficiency plays a fundamental role in fostering effective intercultural communication [4]. In contemporary society, the ability to communicate in multiple languages is a key factor in building intercultural competence, facilitating understanding across linguistic and cultural divides. Language functions as a bridge for communication, comprehension, knowledge exchange, and cultural assimilation, transcending

territorial, ethnic, historical, and religious barriers. It holds an integrative function in intercultural interactions, encompassing both interpersonal and sociocultural dimensions [5].

Studying foreign languages provides valuable insights into different cultures, reinforcing the intrinsic connection between language acquisition and intercultural communication. Research has consistently shown that incorporating digital platforms into conventional foreign language instruction enhances student learning outcomes [6]. Several factors should be considered in this context, as digital platforms not only enhance all language skills—especially listening and speaking—but also promote a more positive attitude toward learning a foreign language while fostering intercultural competence [7].

Establishing effective intercultural communication has become increasingly important, particularly in an era where direct communication across borders is facilitated by digital technologies. The world is more interconnected and mobile than ever before [8]. This necessitates a comprehensive understanding of both the languages being studied and the cultural, historical, and social contexts of different nations. As Ceramella [9] points out, successful cross-cultural interactions require a degree of familiarity with the cultural frameworks of interlocutors.

In this context, fostering intercultural competence and language learning is essential for enhancing global harmony, mutual understanding, and empathy. The rapid advancement of information and communication technologies has introduced innovative platforms that facilitate both language acquisition and intercultural understanding. Numerous studies have explored the benefits of digital platforms for language learning, emphasizing their ability to improve learners' proficiency across various linguistic skills [10]. By incorporating technology into language education, the learning process becomes more engaging and motivating for students. Digital tools provide immediate access to diverse sources of information, enabling more dynamic and interactive learning experiences [11]. This is particularly relevant for younger learners, who find mobile devices and applications more stimulating compared to traditional textbooks and audio CDs. Research indicates that nearly one-third of children aged eight to ten already own a mobile phone and actively use language-learning applications [12].

The foundation of digital language-learning platforms is based on interactive and immersive approaches to foreign language acquisition. These platforms represent a modern pedagogical innovation aimed at enhancing



both linguistic competence and socio-emotional skills. Learning through digital environments is characterized as a transformative process that fosters cognitive and behavioural changes shaped by social interactions, experiences, and cultural practices [13]. Moreover, digital platforms play a crucial role in fostering intercultural competence. Research indicates that collaborative learning through these platforms serves as an effective means of breaking down stereotypes, promoting equal opportunities, and developing key intercultural skills such as communication, teamwork, tolerance, adaptability, and critical thinking [14].

Computer-mediated communication has been shown to enhance intercultural competence by providing learners with authentic opportunities to interact with individuals from different cultural backgrounds [15]. Online platforms enable learners to engage in meaningful intercultural exchanges, helping them develop a deeper understanding of different cultural perspectives [16]. A study [17] confirms that these platforms facilitate cross-cultural communication by exposing learners to diverse cultural traditions, customs, and societal values, thus broadening their global awareness. Furthermore, online intercultural communication platforms play a pivotal role in fostering intercultural dialogue, allowing learners to engage in meaningful discussions that enhance their cultural sensitivity and cross-cultural awareness [18].

3. DIGITAL PLATFORMS FOSTERING INTERCULTURAL COMPETENCE

3.1. DIGITAL PLATFORMS FOSTER INTERCULTURAL COMPETENCE IN FRENCH LANGUAGE CLASSES

TV5 Monde [19] is a multimedia platform dedicated to promoting Francophone culture and fostering intercultural understanding. It encompasses the website apprendre.tv5monde.com [20] and the free educational mobile application *Apprendre le français*. The app serves as a streamlined version of the website, offering learners quick and convenient access to exercises designed to enhance both their language proficiency and cultural awareness, regardless of time or location.

The platform features a rich collection of videos on diverse and authentic topics, each accompanied by comprehension exercises adapted to different proficiency levels, from A1 (beginner) to B2 (advanced). These videos expose learners to a variety of cultural, social, and historical aspects of the French-speaking world, reinforcing not only their linguistic abilities but also their

intercultural competence. Additionally, the platform includes a unique "Memorize vocabulary" section with thematic word lists (e.g., animals, housing, health) and interactive tools like quizzes, flashcards, and listening activities, enabling learners to expand their vocabulary within meaningful cultural contexts.

While users can access the app without an account, registering allows them to track their progress and create a personalized collection of favourite activities, available both on the app and the website. Teachers can also assign activities as homework, encouraging students to engage with authentic French-language materials in and outside the classroom. Each activity consists of a short video, often highlighting current events, social issues, or everyday life in francophone countries, followed by four comprehension exercises. These exercises take various forms, such as true/false questions, sequencing tasks, and multiple-choice quizzes, helping learners to develop both linguistic proficiency and a deeper understanding of francophone cultures. However, as the exercise structure remains relatively consistent, it may feel repetitive over time.

For complete beginners, the full French-language interface might pose challenges without additional support. Nevertheless, TV5 Monde learning program is not intended to replace formal instruction, but rather to complement it by offering authentic exposure to francophone culture. The app excels in this role thanks to its user-friendly design and auto-corrected exercises, which provide instant feedback and encourage independent learning.

Both the app and website include features that support comprehension and cultural immersion, such as adjustable playback speed, optional French subtitles, and full transcriptions. Overall, the TV5 Monde platform serves as a valuable tool for learners aiming to enhance their linguistic skills while gaining intercultural competence and a deeper appreciation of francophone cultures. It also helps teachers integrate authentic, ready-to-use activities that go beyond language learning to explore the richness and diversity of the francophone world.

Podcast Français facile [21] is a well-established online platform designed to facilitate French language acquisition, particularly for beginner and intermediate learners. It provides free access to a wide range of resources, including audio lessons, transcripts, and grammar explanations. By integrating practical conversations, phonetic training, and insights into French and francophone cultures, this platform not only strengthens linguistic proficiency but also fosters intercultural competence.



Over time, *Podcast Français facile* evolved into a comprehensive online learning resource, open to both learners and educators. Its user-friendly interface features an extensive collection of hyperlinked materials, facilitating seamless navigation. The platform hosts more than 500 texts and activities, covering diverse linguistic and cultural topics, including numbers, dialogues, grammar exercises, dictations, poetry readings, and adapted folktales. Each resource is structured around an audio file paired with a printable exercise sheet, which can be used independently or in a classroom setting. To enhance accessibility, earlier flash-based flashcards were later converted into video animations available on YouTube, ensuring compatibility with mobile devices.

Podcast Français Facile proves to be an effective pedagogical tool due to its comprehensive approach to language learning. It provides a wide range of language activities that cater to learners from beginner to advanced proficiency levels, allowing for gradual and structured progression. Each lesson follows a well-organized format, combining theoretical explanations, MP3 audio files that enhance listening comprehension, and exercises that develop productive language skills. This structure enables learners to engage with clearly defined topics in a systematic manner. Additionally, the platform offers flexibility by allowing users to practice listening comprehension offline through printable transcripts and worksheets, making it a convenient resource for independent study. As an open-access platform, it ensures universal availability, removing financial barriers to language learning.

Beyond linguistic development, *Podcast Français Facile* plays a crucial role in fostering intercultural awareness by exposing learners to authentic speech patterns, cultural references, and real-life communicative contexts. Through the inclusion of descriptions of holidays, customs, traditions, folktales, poetry, and everyday dialogues, learners engage with various aspects of Francophone culture, deepening their understanding of cultural nuances, traditions, and communication styles. This immersion not only enhances language proficiency but also cultivates a broader intercultural perspective, equipping learners with the skills necessary to navigate diverse cultural interactions. By integrating both linguistic and cultural dimensions, *Podcast Français Facile* serves as a valuable resource for learners seeking to develop both their command of the French language and their knowledge of Francophone cultures.

3.2. DIGITAL PLATFORMS FOSTERING INTERCULTURAL COMPETENCE IN SPANISH LANGUAGE CLASSES

Ver-taal [22] is a multimedia platform designed for learning the Spanish language and promoting Hispanic culture. It significantly enhances language proficiency while also fostering intercultural competence. The platform serves as an excellent tool for both language acquisition and cultural immersion, offering a vast array of vocabulary and grammar exercises, as well as an extensive collection of multimedia content related to Hispanic culture. It is an ideal supplementary resource for classroom activities and an excellent foundation for homework assignments.

The platform offers a diverse range of content divided into three sections: vocabulary, grammar, and culture. The vocabulary section is structured into three proficiency levels—beginner (*principiantes*), intermediate (*intermedio*), and advanced (*avanzado*). It includes exercises for learning thematic vocabulary (e.g., *La casa*, *la ropa*, *la familia*), as well as interactive activities such as memory games and the classic "hangman" game. A particularly valuable segment for developing intercultural competence is *Vocabulario práctico* (Practical Vocabulary), which provides commonly used phrases in different contexts, such as greetings in formal and informal settings, conversations in restaurants and stores, and interactions at the dentist's office. Many of these expressions are accompanied by audio recordings. Additionally, this section includes exercises on idiomatic expressions, proverbs and sayings (*proverbios y refranes*).

The grammar section offers a wide range of exercises covering basic structures such as articles, pronouns, noun gender and number, and verb tenses. A key feature for enhancing intercultural understanding is the use of adapted authentic texts from Spanish newspapers like *El País* and *El Mundo*, as well as excerpts from reports produced by Spanish National Television (TVE). These resources provide a contextualized approach to grammar learning, reinforcing both linguistic and cultural comprehension.

The culture section of the platform is its most valuable feature when it comes to intercultural competence. It is divided into five categories: movie excerpts (*Trailers*), song lyrics (*Canciones*), advertisements (*Anuncios*), news reports (*Reportajes*), and general culture (*Cultura*). The collection of movie clips is extensive, featuring dialogues from over fifty films produced in the past thirty years in Spanish-speaking countries. This allows users to familiarize themselves with different dialects and regional vocabulary variations. The platform



includes works by directors from Spain, Mexico, Argentina, Colombia, and Chile, providing exposure to a diverse linguistic landscape. The multimedia approach supports the development of multiple language skills—reading, listening, and comprehension—through interactive exercises that provide instant feedback and error correction.

Similarly, the song lyrics section follows an interactive format, allowing students to engage with music by artists across the Spanish-speaking world. The collection includes both contemporary tracks and classics from the mid-20th century, offering cultural and linguistic insights through music.

Ver-taal is a free and user-friendly digital platform that does not require account registration or collect personal data through cookies. It is designed to facilitate Spanish language acquisition while enriching cultural awareness, making it an invaluable tool for developing intercultural competence. Additionally, it offers resources for learning other languages, including French, Dutch, English, Hungarian, and Catalan. The only notable drawback of the platform is its somewhat outdated user interface, but this does not significantly impact its effectiveness in language learning and cultural exploration.

3.3. DIGITAL PLATFORMS FOSTERING INTERCULTURAL COMPETENCE IN ENGLISH LANGUAGE CLASSES

British Council Learn English [23] is a digital platform that promotes intercultural competence, raises awareness of cultural differences, and enables students to explore diverse cultures worldwide in an engaging manner. More specifically, it introduces learners to the Anglophone world through stories, video/audio recordings, and interactive exercises. *Learn English* was developed by the British Council, the United Kingdom's leading organization dedicated to promoting English as a global language and providing support to both learners and educators.

The platform features drop-down menus that allow students to select their language level (ranging from A1 to C1, according to the Common European Framework of Reference for Languages) and the specific language skill they wish to focus on (listening, reading, speaking, or writing). Additionally, dedicated sections are available for grammar and vocabulary, offering a wide range of diverse interactive activities. One of the platform's key advantages is the immediate feedback provided to learners.

It is also worth noting that the platform includes sections for both General English and Business English. Beyond its outstanding potential for enhancing English language proficiency and developing digital literacy, the

Learn English platform serves as an exceptional tool for familiarizing students with the culture, customs, traditions, and heritage of Anglophone countries. Within the General English section, there are subsections such as Video Zone, Audio Zone, Magazine Zone, and Story Zone. These components offer a captivating insight into the culture (and, to a lesser extent, literature) of English-speaking countries, with a particular emphasis on the United Kingdom.

Within the Video Zone and Audio Zone, students can improve their comprehension of English speakers with various accents, as well as their understanding of colloquial and idiomatic language in context. Lessons such as A Study Trip to the UK, Coventry City F.C., Living in London, Music in Manchester, Saint Patrick's Day, Black Friday and Buy Nothing Day, Thanksgiving, Glastonbury Festival, and National Tea Day contribute to a deeper understanding of the culture of Anglophone countries, help overcome cultural barriers, and foster the development of intercultural competence.

To develop confidence in using English in social situations and gain deeper insights into various aspects of British life and culture—including art, history, sports, and music—students can explore the engaging video series Britain is GREAT. Additionally, there is a dedicated set of Shakespeare videos designed for upper-intermediate learners. The World on the Street video series, which features clips of two individuals exploring different locations across the UK, followed by a teacher's explanations of the language used, has proven particularly effective in helping learners overcome both linguistic and cultural barriers. Moreover, this series serves as a strong motivational tool, encouraging EFL students to persist in their language-learning journey.

The BBC Learning English [24] platform is one of the most widely recognized online sites that help people learn English through various multimedia content such as videos, audio, and interactive exercises. Designed and developed by the British Broadcasting Corporation (BBC) the platform is completely free of charge and offers materials for learners at all levels. The platform is divided into different sections that focus on key aspects of language learning: Courses, Grammar and Vocabulary, Pronunciation, News and Current Affairs. In addition to the above sections, there are also various subsections, including Drama, Business, Teachers, Podcast, Quizzes, etc. 'Grammar and Vocabulary' section includes grammar lessons, everyday words and phrases, and practice exercises, making learning easy to follow. The lessons in the 'Pronunciation' section use audio and video resources to enable users to improve their listen-



ing and pronunciation skills by having them repeat what native speakers are saying. 'News and Current Affairs' section provides short news items and analyses of the current affairs thereby helping learners gain an insight into real-life English in context. The 'Business English' section is intended for those who wish to improve their English in the workplace setting through lessons on business vocabulary and communication, and workplace etiquette. The section includes workplace scenarios such as meetings, negotiating, presenting, and email writing. Series like English at Work and Business Daily provide interesting conversations in real business settings. Each lesson is accompanied by transcripts, key phrases, and practice exercises. The 'Drama' section uses the art of storytelling to help learners improve their English. It provides serialized audio dramas with engaging plots, professional narration, and authentic dialogues. Some of the popular dramas include *The White Elephant*, *A Christmas Carol*, and *The Race*. Through comprehension exercises, vocabulary explanations, and transcripts learners are enabled to improve their listening skills, pronunciation, and vocabulary in an enjoyable way.

BBC Learning English is a globally accessible platform with easy navigation through user-friendly layout. It offers multimedia format, i.e., reading, listening, and interactive exercises to cover different learning styles. Mobile applications and social media are its additional facilities for learning on the go. Through real-world dialogues, cultural insights, and global news stories, learners not only improve their English proficiency but also develop an understanding of different cultures, customs, and communication styles. Testimonials and research evidence prove it has helped millions learn English, and it is a significant force in global language learning.

4. CONCLUSION

In today's increasingly interconnected world, fostering intercultural competence has become an essential goal of language education. This paper explored how digital platforms used in foreign language classes contribute to the development of intercultural competence in language classrooms. These platforms offer learners unique opportunities to immerse themselves in authentic cultural contexts, engage with diverse language resources, and enhance their understanding of the cultural nuances that shape communication. By integrating linguistic and cultural content, these platforms not only support language acquisition but also enable learners to navigate the complexities of cross-cultural interactions.

Platforms for learning French, Spanish, and English described in the paper illustrate how digital tools can be effectively utilized to expose students to a broad range of authentic materials, from everyday dialogues and folktales to poetry and cultural discussions. These resources facilitate the development of both receptive and productive language skills, while also promoting a deeper appreciation of the cultural dimensions of the language. In doing so, learners are better prepared to engage with the international world, equipped with the linguistic proficiency and cultural sensitivity necessary for meaningful intercultural exchanges.

The insights from this paper have several important pedagogical implications. First, language educators should embrace digital platforms as valuable supplementary tools to enhance traditional classroom learning. These platforms offer flexible and accessible resources that allow learners to independently explore the cultures of the language being studied, fostering a more autonomous and self-directed approach to language learning. By incorporating these platforms into their teaching practices, educators can provide students with authentic cultural content that enriches their language proficiency and intercultural understanding.

Second, teachers should consider the integration of both linguistic and cultural learning objectives in their curriculum design. The platforms mentioned in French, Spanish, and English exemplify how digital tools can bridge the gap between language instruction and cultural exploration, providing learners with a holistic language learning experience. Educators can curate digital content that reflects global diversity, including not only linguistic exercises but also materials that introduce cultural traditions, customs, and everyday communication practices.

Additionally, while these digital platforms offer valuable opportunities for intercultural competence development, educators must be mindful of potential challenges, such as the lack of formal assessment tools on these platforms. To address this, teachers should complement the use of digital resources with in-class activities that assess students' progress in both linguistic and intercultural areas. This could include guided discussions, group projects, or reflective exercises that allow learners to apply the cultural insights gained through digital platforms to real-world contexts.

Finally, the importance of digital literacy in language education cannot be overstated. As digital platforms become an integral part of language learning, educators must ensure that students are equipped with the skills to navigate these resources effectively.



This includes not only the technical ability to use the platforms but also the critical skills to analyse and interpret cultural content, thus deepening their intercultural understanding.

Digital platforms offer powerful tools for fostering intercultural competence in language classrooms. By thoughtfully integrating these resources into the curriculum, educators can enhance both the linguistic and cultural dimensions of language learning, preparing students for the challenges and opportunities of an increasingly globalized world.

REFERENCES

- [1] K. Abisheva, K. Nagimzhanova, R. Aykenova, A. Kapanova, Z. Koldasbayeva and A. Kanapyanova, "Cross-cultural competence of communicators as a way to create a positive eco-informational environment," *Space and Culture*, vol. 7, no. 2, pp. 76-78, 2019.
- [2] D. S. Petrović, *Umetnost komuniciranja - teorijski i praktični aspekti*, Beograd: Clio, 2019.
- [3] K. Cushner and J. Mahon, "The intercultural speaker and the pedagogy of foreign language education," pp. 304 - 320, 2009.
- [4] H. Y. Al Jarrah, "Six thinking hats: An analysis of the skill level of Jordanian vocational education teachers and the extent of skill application," *Space and Culture*, vol. 7, no. 1, pp. 170-185, 2019.
- [5] S. G. Ter-Minasova, *Language and intercultural communication*, Moscow: SLOVO, 2000.
- [6] H. S. Kashan and A. F. AbuSeileek, "The Effect of Using British Council BBC Learning English Application on the Tenth Grade EFL Students' Listening and Speaking Skills and their Attitudes toward it," *Journal of Applied Linguistics and Language Research*, vol. 10, no. 2, pp. 39-62, 2023.
- [7] A. H. Maulana, A. Syah Putra and A. Rohim, "Optimizing English Language Training Through British Council Learn English Kids in Banthamaprao School Thailand," *LOSARI: Jurnal Pengabdian Kepada Masyarakat*, vol. 6, no. 2, pp. 229-235, 2024.
- [8] K. Korhonen, "Developing Intercultural Competence as Part of Professional Qualifications. A Training Experiment," *Journal of Intercultural Communication*, vol. 4, no. 1, pp. 1-6, 2004.
- [9] N. Ceramella, "Intercultural Communication in Different Cultures," *Journal of Intercultural Communication*, vol. 14, pp. 15-24, 2008.
- [10] K. Mullamaa, "ICT in Language Learning-Benefits and Methodological Implications," *International Education Studies*, vol. 3, no. 1, pp. 38-44, 2010.
- [11] T. Gajić and N. Maenza, "Korišćenje mobilnih aplikacija u nastavnom procesu učenja stranih jezika na tercijarnom obrazovnom nivou," *Inovacije u nastavi*, vol. 32, no. 4, pp. 133-142, 2019.
- [12] V. J. Rideout, U. G. Foehr and D. F. Roberts, "Generation M2: Media in the lives of 8-18 year-olds," January 2010. [Online]. Available: <http://kff.org/other/poll-finding/report-generation-m2-media-in-the-lives/>. [Accessed 2 March 2025].
- [13] N. Suzić, P. Stojaković, M. Ilić, D. Branković, S. Milijević and D. Krneta, *Interaktivno učenje*, Banja Luka: Ministarstvo prosvete Republike Srpske i UNICEF Kancelarija u Banjoj Luci, 1999.
- [14] S. Romić, "Isti u različitim," *Školske novine*, vol. 62, no. 21, pp. 10-11, 2011.
- [15] G. Chen and W. Starosta, "The Development and Validation of the Intercultural Sensitivity Scale," *Human Communication*, vol. 3, pp. 1-15, 2000.
- [16] C. Kramsch and S. L. Thorne, "Foreign language learning as global communicative practice," in *Globalization and language teaching*, London, Routledge, 2002.
- [17] L. Lee, "Focus-on-form through collaborative scaffolding in expert-to-novice online interaction," *Language Learning and Technology*, vol. 12, no. 3, pp. 53-72, 2008.
- [18] J. Kim and J. R. Lee, "The Facebook Paths to Happiness: Effects of the Number of Facebook Friends and Self-Presentation on Subjective Well-Being," *Cyberpsychology Behavior and Social Networking*, vol. 14, pp. 359-364, 2011.
- [19] "TV5Monde : TV Internationale francophone : Info, Jeux, Programme," TV5Monde, [Online]. Available: <https://www.tv5monde.com>. [Accessed February 2025].
- [20] "TV5Monde : Apprendre le français," TV5Monde, [Online]. Available: <https://apprendre.tv5monde.com/fr>. [Accessed February 2025].
- [21] "Podcast Français facile," [Online]. Available: <http://www.podcastfrancaisfacile.com/>. [Accessed February 2025].
- [22] "Ver-taal," [Online]. Available: <https://www.ver-taal.com/>. [Accessed February 2025].
- [23] "Learn English," British Council, [Online]. Available: <https://learnenglish.britishcouncil.org/>. [Accessed February 2025].
- [24] "Learn English with BBC Learning English," BBC, [Online]. Available: BBC Learning English - Learn English with BBC Learning English - Homepage. [Accessed February 2025].



CHALLENGES IN USING ACTIVE LEARNING IT STRATEGIES IN FOREIGN LANGUAGE TEACHING FOR HIGHER EDUCATION

Milica Pančevac*,
[0009-0007-6011-3014]

Ana Tripković
[0009-0009-4541-7380]

Singidunum University,
Belgrade, Serbia

Abstract:

This paper examines the difficulties of applying IT strategies for active learning in foreign language instruction at higher education institutions. By integrating artificial intelligence (AI) and information and communication technologies (ICT) with active learning, which is already recognized for improving student engagement and encouraging critical thinking, its importance has been continuously growing. These technologies provide enhanced accessibility, interactive content, and customized learning experiences. Their use in teaching foreign languages, however, brings several difficulties.

The attitudes of foreign language teachers in higher education were explored through a survey indicating that they are aware of both benefits and challenges. Digital literacy among teachers and students was found to be one of the main issues. Students may find it challenging to adapt to new digital learning environments, while educators may encounter difficulties in effectively integrating technology without adequate training. Teachers are further burdened by the need for ongoing professional development due to the quick evolution of IT. The accessibility of technological resources is another important concern. Financial limitations frequently prevent institutions from purchasing and maintaining the required infrastructure. Furthermore, survey respondents stated that technical issues are the most common source of disruption.

The review of relevant literature, along with the responses provided by teaching professionals, suggests solutions such as focused professional growth and calculated investments in IT infrastructure. Higher education institutions will be able to employ IT strategies in active learning for foreign language instruction more successfully if these problems are resolved.

Keywords:

Active Learning, IT Strategies, Higher Education, Challenges.

INTRODUCTION

The implementation of IT strategies for active learning in higher education, in particular in teaching foreign languages, presents several significant challenges. The main problem is the existence of various levels of technological competence among students and instructors, which can hinder the effective use of digital tools (Bao, 2020). Since most institutions rely on online platforms, comprehensive training and support are essential to ensure that educators can integrate these tools effectively in their pedagogical practices (Jovanović 2015).

Correspondence:

Milica Pančevac

e-mail:

mpancevac@singidunum.ac.rs





Additionally, the lack of involvement and motivation among students can represent a significant barrier to active learning strategies. The use of game-based technologies, such as Kahoot or Quizzes have been shown to improve the involvement of students and facilitate a more interactive learning environment (Plump & Larosa, 2017). However, the challenge which remains is in ensuring that these strategies are applied consistently (Rapanta et al., 2021). In addition, the flipped classroom model has been proposed as an effective means to promote active learning in linguistic education, but its success largely depends on the preparation of students and the desire to engage with pre-class materials (Foldnes, 2016). In order to deal with these challenges, higher education institutions must give priority to professional development for educators and encourage a culture that enhances collaboration and innovative teaching practices thereby improving the results of learning foreign languages (Niemi et al., 2016).

2. REVIEW OF LITERATURE

The theoretical framework of active learning relies on constructivism and socio-constructivism which promote the role of students as active participants in the knowledge acquisition process [1]. Constructivism indicates that learners develop their own understanding through interactions with their environment and prior knowledge [2]. Eminent social constructivist Lev Vygotsky [3], described the concept zone of proximal development (ZPD). This is the field where learning activities should be centered, between what the learner can accomplish independently and what the learner can complete with expert guidance. ZPD basically refers to eliciting students' ability to acquire knowledge on their own while also examining and suggesting that students are capable of acquiring more easily more complex concepts with the help of peers and professors. According to constructivists, knowledge is fundamentally subjective and is created by our shared perceptions and customs [4]. On the other hand, social constructivists claim that social interaction and conversation are the best ways for this process to occur because they let us compare and contrast our own understandings with those of others.

3. THE IMPORTANCE AND BENEFITS OF ACTIVE LEARNING IN FOREIGN LANGUAGE TEACHING

The most important methods which motivate students to think critically are those that use problem-solving practices and collaboration [5]. Some of the benefits of active learning are [6]:

- **Improved Understanding and Retention of Information:** Active learning strategies, such as collaborative tasks, peer teaching, and problem-solving activities, engage learners both cognitively and emotionally leading to enhanced memory retention, unlike traditional passive methods [7]. Interactive activities deepen the understanding of complex linguistic concepts [8]. Moreover, active learning activities facilitate the process of reflection and students can apply learned concepts to real-life situations with greater ease [9]. These insights illustrate that active learning strategies support the process of retention.
- **Development of Critical Thinking:** It is discussed that active learning actually facilitates the process of analyzing, synthesizing and assessing information and in this way, students are not just passively absorbing information [10], they are involved in creating them. Peer interactions are promoted by incorporating active learning strategies such as discussions, debates, project-based learning, peer teaching which is of crucial importance for critical thinking development [11].
- **Increased Motivation and Engagement:** Active learning strategies, especially when combined with collaborative learning, have been shown to significantly boost students' motivation and academic performance. One of the studies shows that a repeated measures experimental design across multiple engineering programs implies 14% improvement in student performance after implementing active learning techniques [12]. In addition, it was assessed that the impact of active learning on student motivation in a process control course resulted in increased motivation and improved performance for 40% of students [13]. Active learning also nurtures students' autonomy [14], giving them greater control over their learning process.



- **Enhanced Social and Communication Skills:** Many studies confirm that active learning methods not only improve academic outcomes but also foster critical thinking, social skills and communication skills. According to Yaacob and Mahmud [15] who incorporated active learning with social skills training in higher education and applied both qualitative and quantitative methods to assess students' social skills development, it was revealed that students who engaged in group discussions and collaborative tasks improved not only their communication but also interpersonal abilities. Agbatogun [16] explored the impact of active learning on second-language learners' communicative competence, using clicker technology [17] and communicative tasks. Surveys and language assessments indicated that active learning helped students improve their ability to communicate in target language.

4. THE INTEGRATION OF AI AND INFORMATION AND COMMUNICATION TECHNOLOGIES (ICT) INTO ACTIVE LEARNING STRATEGIES IN FOREIGN LANGUAGE TEACHING AT HIGHER EDUCATION INSTITUTIONS

4.1. CONTEXT

The implementation of artificial intelligence (AI) and information and communication technologies (ICT) into active learning strategies represents remarkable progress in foreign language teaching at higher education institutions. This approach enables personalized learning, boosts student engagement, and improves learning outcomes [18].

- **Personalization and Adaptive learning** [19] involve using technology in order to adjust materials, lessons, handouts, presentations etc. to individual students' needs taking into account the students' weaknesses and potentials. Using AI, lecturers can provide customized handouts, papers, assignments, and feedback. As a result, students can progressively advance according to their abilities. AI makes it possible to design learning programs that adapt to each student's unique requirements and learning styles [18].
- **Interactivity, Student Engagement and Automated Feedback and Assessment:** The creation of interactive platforms that promote active student participation is simplified by the integration

of AI and ICT in foreign language instruction. These types of AI-driven tools also enhance students' interactive experiences, enabling more authentic and engaging language use in real-life situations [20]. Studies have shown that automated feedback systems dramatically boost student engagement and motivation [21]. AI-based tests, like speech recognition and grammar checkers, can provide comprehensive information about students' language skills, enabling teachers to address the challenges that need to be resolved more successfully. These tools assist teachers to improve their methods of instruction and to promoting learning [20].

5. CHALLENGES IN THE IMPLEMENTATION OF IT STRATEGIES FOR ACTIVE LEARNING IN HIGHER EDUCATION IN THE CONTEXT OF FOREIGN LANGUAGE TEACHING

- **Lack of Technological Infrastructure:** One of the biggest challenges to the effective application of IT-based active learning strategies, particularly in the context of teaching foreign languages, is the technological infrastructure within institutions. Lower funding is the main reason for this [22]. Moreover, it implies the fact that acquisition and maintenance of advanced IT resources can be financially challenging for some universities.
- **Varying Levels of Digital Literacy:** Digital literacy is particularly noticeable in foreign language courses, where encouraging active learning requires the use of interactive applications, online collaboration platforms, and digital language resources. On the one hand this can influence students' engagement and instructional efficacy [22]. Moreover, disparities in digital proficiency hinder teachers from effectively integrating IT strategies into their lessons [23].
- **Instructor Training:** Many educators find it difficult to adapt their traditional teaching methods to incorporate digital resources, which can limit the effectiveness of active learning approaches [24]. The lack of tailored professional development, lecturers' digital literacy can hinder the implementation of online teaching strategies, ultimately impacting the quality of instruction and student participation [25]. To resolve this problem, universities must invest in training programs that provide educators with the skills and knowledge required to use IT tools effectively in the foreign language classroom.



- **Student Motivation:** If students do not comprehend how technology enhances their learning, they are less likely to participate in IT-based activities [26]. In order to overcome this challenge, teachers need to show students the benefits of IT-driven learning and design activities that encourage motivation and active participation.
- **Language Barriers:** Various digital tools are made in English, which can be challenging for non-native speakers to use. Chapelle [27] argues that language barriers in digital environments can cause frustration and lower levels of motivation, especially when neither lecturers nor students are competent and experienced in using the IT tools. These barriers can also lead to resentment and lower motivation.

6. RESEARCH

6.1. CONTEXT

The research was conducted with 40 foreign language teachers who work in higher education with the aim of getting their attitudes towards using active learning IT strategies in their work and identifying possible obstacles they might be facing. They were sent an online survey consisting of 15 questions, six of which were close-ended multiple choice questions, two checkbox questions allowing multiple answers and adding additional information, four Likert scale questions (ranging from 1 to 5) and three open-ended questions.

To evaluate the data from this small-scale study, a mixed-method approach was used, combining quantitative descriptive statistical analysis of close-ended

questions—performed using IBM SPSS Statistics software (version 26)—with a qualitative analysis of open-ended responses.

6.2. QUANTITATIVE RESEARCH

The sample consisted of 35 female (87.5%) and 5 male (12.5%) participants, with the majority (25 or 62.5%) being language teachers and a significant number of associate professors (8 or 20%). In total 29 teachers (72.5%) have more than 11 years of teaching experience which is an interesting fact, since 18 (45%) participants have more than 11 years of experience in higher education. Overall, it can be concluded that the sample consists of highly experienced professionals predominantly teaching English language (34 or 85%).

Pearson correlation coefficient does not show really strong statistical correlation between any of the given variables and the only moderately strong correlation visible is the one between confidence in using IT strategies and familiarity using active learning IT strategies. A correlation of 0.724 suggests that as confidence in using IT strategies increases, familiarity with IT Active Learning Strategies also tends to increase. Since $p < 0.01$, the correlation is statistically highly significant.

Table 1 presents descriptive statistics for Likert scale questions from the survey. The mean values of 3.80 for most questions and one 3.73 indicate that most participants feel confident and familiar with the given strategies. However, standard deviation suggests that most participants have different opinions about the significance of institutional support.

Table 1. Descriptive statistics for questions about active learning IT strategies

Question	Minimum	Maximum	Mean value	Standard deviation
How would you assess your familiarity with IT active learning strategies?	2	5	3.80	0.883
How much is institutional support significant for successful implementation of active learning strategies?	2	5	3.80	1.018
How frequently do you integrate IT-based active learning methods into your teaching?	1	5	3.73	0.987
How confident do you feel using IT strategies?	1	5	3.80	0.939



Language teachers in higher education claim to be using all kinds of active learning strategies, predominantly choosing interactive quizzes and AI-driven tools (32=80% and 27=67.5% respectively). The most challenging thing in using these activities is technical issues (26 participants = 65%) and an interesting finding is that 16 teachers (40%) claim that lack of training can be an obstacle, even though a vast majority is highly experienced.

Overall statistical data emphasize positive perception of active learning IT activities. This is further highlighted in the answers to the question about the fact that these activities enhance students' engagement and participation, where 85.2% (33 participants) agreed with the statement, whereas only 7 of them were not sure about it and not a single participant believes this is not true.

7. QUALITATIVE RESEARCH

For the qualitative part of the research, participants were asked some open-ended questions and the answers show a lot of interesting insights into the major obstacles and opinions of teachers towards active learning IT strategies. Given that the major obstacle for most respondents was technical issues, there are numerous suggestions on how to deal with them. One of the subjects wrote "Plan ahead: Select IT tools that are user-friendly and require minimal setup." Then it was suggested to check all the tools and the internet connection before the lesson, prepare printed backup material or another activity, practice using the tools before applying them in the lessons, learning about them from workshops, YouTube videos or self-study materials. Thus, it can be concluded that there is always a backup plan, often stemming from experience and in case that doesn't work, six colleagues refer to IT support at universities.

When it comes to pedagogical challenges that might be encountered, there is a range of interesting responses. Five respondents claim they almost never face such challenges. In contrast there are statements that sheer variety of options can be overwhelming, which is connected to the lack of time for working on new tools. Additionally, even though students get used to it over time, they are also the ones who might feel overloaded with complex programs. Varying levels of digital literacy and language proficiency in students could also cause issues but "keeping students actively involved and avoiding passive consumption" is another challenge teachers might face. Even though students perceive these activities as fun, but they sometimes do not see the learning goals, or, as

one participant stated, they lose interest in traditional textbooks and it gets increasingly difficult to motivate them. Moreover, engaging students who are not confident is identified as an issue, but another response mentions that sometimes, in anonymous activities, certain students do not even participate.

Teachers may spend too much time on preparing such activities, for this reason one of the participants claims that they do not use these activities even though they are aware of all the pros and cons of using them, simply because teachers have to do extra work due to small salaries and do not have time to learn more things about these strategies.

Finally, as suggestions or additional strategies respondents claim that some flipped classroom activities might be efficient where students have to do certain tasks at home, before coming to class and then engage in activities. Another suggestion is to use them as a treat, as a reward, rather than incorporating it in a lot of activities. Similarly, it is advised to provide additional points.

An amazing array of resources was mentioned within the survey and here is the list: Hot potatoes app, bamboozle, WordWall, Canva, Storyboard, Vocaroo, Google docs, AI powered language tool, Socratic Seminar through Google classroom and Microsoft Teams. The fact that a lot of teachers specifically mentioned tools they had been using in their teaching shows the importance and prevalence of such activities in everyday lessons.

8. CONCLUSION

The research emphasizes both challenges and potential of including active learning IT strategies for teaching foreign languages at the university level. While new digital tools bring more positive energy and clearly show enhanced language acquisition, they also pose huge issues when it comes to their integration into standard curricula. Key issues such as varying levels of digital literacy among students and teachers, the lack of adequate institutional support, and the financial constraints on universities hinder the widespread and effective use of these technologies. Furthermore, it is obvious from the responses that regardless of their experience, teachers do expect more adequate training. Addressing the issues raised by them, particularly about additional training, IT support, lack of time, can help successful integration of active learning IT strategies into regular foreign language curricula at universities.



This paper outlines the main problems and advantages from the teachers' perspective, providing only one side of the topic. The suggestions for further research would include the students' view and could focus on specific comparison of results achieved by incorporating active learning IT strategies. Integrating all the suggestions into everyday teaching activities could lead to better language acquisition, but it could also prepare students for the requirements of the new digital age.

REFERENCES

- [1] C. English, "Guided support for writing assessment literacy (GSWAL)," Cambridge University Press & Assessment, [Online]. Available: <https://www.cambridge-community.org.uk/professional-development/gswal/index.html>. [Accessed 28 March 2024].
- [2] J. Piaget and B. Inhelder, The psychology of the child, London: Basic Books, 2008, pp. 84-92.
- [3] S. L. Vygotsky, Mind in Society The Development of Higher Psychological Processes, London: Harvard University Press, 1978, pp. 87-97.
- [4] W. A. Bates, Teaching in a digital age: Guidelines for designing teaching and learning, Vancouver: Tony Bates Associates Ltd., 2015, p. 60.
- [5] J. Allsop, S. J. Young, E. J. Nelson, J. Piatt and D. Knapp, "Examining the benefits associated with implementing an active learning classroom among undergraduate students," *International Journal of Teaching and Learning in Higher Education*, vol. 32, no. 3, pp. 418-426, 2020.
- [6] E. F. Berkley, P. K. Cross and C. H. Major, Collaborative learning A handbook for college faculty, San Francisco: Jossey-Bass, 2005, pp. 95-126.
- [7] S. Freeman, S. L. Eddie, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt and M. P. Wenderoth, "Active learning increases student performance in science, engineering and mathematics," in *Proceedings of the National Academy of Science of the United States of America*, San Francisco, 2014.
- [8] T. C. M and R. Wylie, "The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes," *Educational Psychologist*, vol. 49, no. 4, pp. 219-243, 2014.
- [9] M. Prince, "Does Active Learning Work? A Review of the Research," *Journal of Engineering Education*, vol. 93, no. 3, pp. 223-231, 2004.
- [10] C. Bonwell and J. Eison, "Active Learning: Creating Excitement in the Classroom," The George Washington University, School of Education and Human Development, Washington DC, 1991.
- [11] D. W. Johnson, R. T. Johnson and K. A. Smith, "Cooperative learning: Improving university instruction by basing practice on validated theory," *Journal on Excellence in College Teaching*, vol. 25, no. 4, pp. 85-118, 2014.
- [12] L. M. Guimarães and R. S. Lima, "Active learning application in engineering education Effect on student performance using repeated measures experimental design," *European Journal of Engineering Education*, vol. 46, no. 5, pp. 813-833, 2021.
- [13] Z. Dadach, "Quantifying the effects of an active learning strategy on the motivation of students," *International Journal of Engineering Education*, vol. 29, no. 4, pp. 904-913, 2013.
- [14] C. A. I. Education, "Active Learning," April 2020. [Online]. Available: <https://www.cambridgeinternational.org/Images/271174-active-learning.pdf>. [Accessed 21 February 2025].
- [15] Y. Yaacob, M. M. Mahmud and W. S. Ching, "The impacts of active learning environment on communication skills, thinking and problem-solving skills, and teamwork skills," in *Proceedings of the 4th International Conference on Information and Education Innovations (ICIEI '19)*, Durham, 2019.
- [16] A. O. Agbatogun, "Developing learners' second language communicative competence through active learning: Clickers or communicative approach?," *Journal of Educational Technology & Society*, vol. 17, no. 2, pp. 257-269, 2014.
- [17] V. Nejković and D. Andrejević, "The Impact of The Use of Clicker Tools in Academic Classroom Environments," *The Journal of Teaching English for Specific and Academic Purposes*, vol. 10, no. 3, pp. 407-420, 2022.
- [18] A. Varsamidou, "Integration of artificial intelligence in foreign language Exploring views and beliefs of Greek foreign language teachers," *International Journal of Language and Literary Studies*, vol. 6, no. 3, pp. 299-320.
- [19] T. Cavanagh, B. Chen, R. A. M. Lachen and J. R. Paradiso, "Constructing a Design Framework and Pedagogical Approach for Adaptive Learning in Higher Education: A Practitioner's Perspective," *International Review of Research in Open and Distributed Learning*, vol. 21, no. 1, pp. 173-197, 2020.
- [20] A. Edmett, H. Crompton, N. Ichaporia and R. Crishton, "Artificial intelligence and English language teaching: Preparing for the future," British Council, London, 2023.
- [21] R. Schiller, J. Fleckenstein, U. Mertens, A. Horbach and J. Meyer, "Understanding the effectiveness of automated feedback: Using process data to uncover the role of behavioral engagement," *Computers & Education*, vol. 223, pp. 2-16, 2024.



- [22] T. Bates, *Teaching in a Digital Age: Guidelines for designing teaching and learning* - 2nd Edition, Victoria: BCcampus, 2015, p. 60.
- [23] M. Healey, A. Flint and K. Harrington, *Engagement through partnership: students as partners in learning and teaching in higher education*, Yorkshire: The higher education Academy, 2014.
- [24] D. W. Johnson, R. T. Johnson and K. A. Smith, "Co-operative learning: Improving university instruction by basing practice on validated theory," *Journal on Excellence in College Teaching*, vol. 25, no. 4, pp. 85-118.
- [25] C. A. Hafner, A. Chik and R. H. Jones, "Digital literacies and language learning," *Language Learning & Technology*, vol. 19, no. 3, pp. 1-7.
- [26] A. G. Negoescu and C. M. Mitrulescu, "Using technology to increase students' motivation for learning a foreign language," in *Knowledge-Based Organization*, Sibiu, 2023.
- [27] C. A. Chapelle, "The interface of technology and language learning: Challenges and opportunities," *The Modern Language Journal*, vol. 94, no. 1, pp. 115-130, 2010.



DISRUPTION IN COMMUNICATION IN HYBRID FOREIGN LANGUAGE TEACHING MODEL

Katarina Nasradin*,
[0000-0002-8855-7174]

Miloš Pupavac,
[0000-0002-6179-3900]

Maja Veljković Michos,
[0000-0003-2931-6554]

Valentina Bošković Marković,
[0000-0001-8873-6706]

Darija Lunić
[0000-0002-1358-4696]

Singidunum University,
Belgrade, Serbia

Abstract:

In the new digital era, within the hybrid teaching model used for foreign languages at Singidunum University, both teachers and students face various challenges: lack of communication between students in the classroom and those attending via Teams, scattered attention of lecturers who teach both in the classroom and to students online, reduced eye contact among interlocutors, etc. Therefore, the main goal of this paper is to analyze disruption in communication in the hybrid foreign language teaching model. For the purpose of this study, the authors have conducted quantitative analysis, using two online questionnaires: one for students, and one for foreign language teachers. The sample includes 125 students and 25 foreign language teachers. According to the research results, along with the number of benefits, the hybrid teaching model also introduces several drawbacks, such as the lack of eye contact, hindered interaction, and reduced non-verbal communication.

Keywords:

Disruption, Communication, Hybrid, Language, Interaction, Classroom, Students.

INTRODUCTION

According to Pauković and Krstinić [1], the hybrid teaching model is a combination of teaching in person and teaching by using digital technology simultaneously. We have already covered this topic in our previous studies, especially by mentioning its advantages. First, it allows those who are accustomed to the traditional way of learning (in person) to follow the course flow. On the other hand, it offers working students the possibility to simultaneously attend lectures from another location, saving them both time and money. This teaching model can use the most contemporary materials from the Internet, which are visible via a projector to those in the classroom, as well as to those attending classes through the Microsoft Teams platform. At Singidunum University, from which we took the sample for our study, the Microsoft Teams platform is used in the hybrid teaching model for all courses. It is fascinating how quickly students adapted to this type of teaching and the use of digital tools. The use of digital tools in a hybrid foreign language teaching model will be further explained in this paper.

Correspondence:

Katarina Nasradin

e-mail:

knasradin@singidunum.ac.rs





A part of the research results is taken from one study [2], with particular focus on the question: *Do you find it easier to learn a language with ICT tools?* According to those results, 112 respondents agreed with this, while only 10 did not. Another study [3] shows that all respondents still prioritize a teacher-led teaching process and that digital tools cannot replace the teacher.

According to Stanković and Stoimenov [4], information and communication technologies are crucial in contemporary society in educating new generations who live in the Internet era. Informatics and computer science have been introduced as a separate course in schools starting from the first grade since, without computer literacy, newer generations would not be strong competitors in the future job market. It is somehow implied that everyone needs to be digitally literate. In this regard, they should be capable of navigating a hybrid environment.

1.1. COMMUNICATION AND COMMUNICATION ISSUES

If we still believe that language learning is accelerated through teachers or mentors, then we need to highlight the importance of communication, which has significantly changed compared to traditional teaching methods. According to the Referral Center for the Methodology and Communication of E-learning [5], communication in education consists of the following components:

- Informational-cognitive,
- Socio-emotional.

The messages that are exchanged in written or spoken form via Teams — with or without the use of a camera — can be verbal and non-verbal (eye contact, voice — tone, intonation, pace, body language — gestures, movements, clothing, the distance between speakers, etc.). Moreover, the communication climate or atmosphere in the classroom where a foreign language is studied is also significant. We call this climate the ‘energy’ of the classroom. Participants should always feel safe, accepted, and relaxed, without fear of making mistakes or speaking in front of a group, and the teacher can always influence this.

When it comes to the socio-emotional components of communication, they are more important than mere knowledge of a foreign language. Speaking is ineffective if the participant cannot navigate social environments or society. Furthermore, emotional intelligence is closely related to social intelligence, which includes the ability to react appropriately in an emotional sense,

to endure unpleasant emotional experiences, to recognize emotions in others, resolve conflicts, and generally adapt within social settings. This competence is much more demanding to develop via Teams than in real-life communication. From a certain perspective, society is indeed in crisis, largely since these competencies are being increasingly neglected because of the influence of digitalization. This leads to alienation, withdrawal, lethargy, and other issues. The educational process should not only focus on sharing knowledge, but it should also include the upbringing and moral development of students and children. According to Klimova [6], a teacher needs to adopt an individualized approach to each student, both through Teams and in the classroom. Being a teacher in modern times is highly challenging. A teacher becomes a tutor who manages teaching processes, who is a mentor, or a guide. This role is far more extensive than it once was, as mentoring now occurs both digitally and in person. In the post pandemic landscape [7], [8] current challenges are confirmed in hybrid learning in the higher education context, especially when it comes to teacher-student interaction, due to the previously mentioned limitations.

If we were to analyze the hybrid teaching model conducted at Singidunum University in the previous few years, with a particular focus on foreign language teaching (the languages included are English, German, Russian, Spanish, Italian, French, Greek, and Chinese), we could highlight the main communication issues. First, eye contact is reduced—both with those on Teams via the camera and with those sitting in the classroom. The teacher simultaneously monitors both methods and divides attention between the two. When we don't look the interlocutor directly in the eyes, our attention tends to scatter more quickly. The second major issue is the teacher's immobility, the inability to move around the classroom, as they often adjust materials on the computer and ensure that those attending via Teams can see them continuously online. This immobility can lead to monotony and makes it challenging for students to watch and listen to the lecturer in the same setting for an extended time. The third problem is sound. Participants attending via Teams cannot always hear those in the classroom who read texts or ask questions clearly. On the other hand, the students who are in the classroom can hear participants on Teams well. The fourth issue is the exchange of non-verbal communication between the teacher and participants, as well as non-verbal communication, socializing, and interaction among participants themselves. In large groups, even if all participants turn on their cameras, the teacher cannot simultaneously



see the facial expressions of all participants. Frequently, participants do not turn on their cameras (because they are calling from work or are unable to), and the teacher sometimes has the impression that they are teaching a computer rather than people. Non-verbal communication, which often conveys much more than verbal communication alone, provides the teacher with feedback on whether they are doing something well or not. For instance, imagine a foreign language teacher dictating a text to their participants. If all cameras are turned off, the teacher will not know if they are dictating too quickly or too slowly, if the dictation is too difficult or too easy, or if they are expressive enough. Communication via the Internet is a new form of communication in which it is even more challenging to meet all participants' needs. This brings us to the fifth and perhaps the main issue: participants' focus and concentration. Feelings of lethargy, boredom, and monotony arise. Digitalization and the use of mobile phones and computers have led to reduced concentration—not only in developing children but also in adults. Constant interruptions and jumping from one content to another have altered our focus and attention. While attending several classes via the Microsoft Teams platform, a significant number of participants struggle to follow the course flow. Their thoughts wander, especially when disruption occurs, as this causes them to lose the thread. Those in the classroom have an advantage, at least in terms of attention. However, even here, problems arise. Most participants follow via the platform, leaving those in the classroom feeling lonely or too much at the center of the teacher's attention. Eventually, the teacher dominates the teaching process, and participants contribute to a lesser extent than ever. In foreign language learning theory, it has always been said that a successful foreign language class is one in which participants speak, reproduce the language, and the teacher remains silent. Nowadays, the teacher dominates the class. Participants find it easier to speak a foreign language with an interlocutor in person because this provides them with greater confidence. As a result, many eventually choose not to attend in person. The sixth issue is the lack of socialization and teamwork. These are factors that satisfy the need for belonging, which, according to Maslow, is a very important need that is being lost in this method.

Regarding technical issues that occasionally occur in the hybrid model (Internet traffic quality, computer problems, power outages, etc.), even though they are rare, when they do happen, the teaching process—and consequently the communication process—completely halts or becomes disrupted. This is most certainly another drawback of the hybrid model.

2. RESEARCH METHODOLOGY AND RESEARCH RESULTS

The main goal of this study was to examine the types and extent of disruption in communication that occur during foreign language teaching in a hybrid model at various faculties and study programs at Singidunum University in Belgrade. Additionally, we aimed to compare the attitudes of foreign language teachers and students by distributing two separate questionnaires with similar questions.

The sample consisted of 125 students (91 female and 33 male students, with one respondent not declaring their gender) from undergraduate and master's programs, and 25 foreign language professors. Students from the following faculties and study programs were included in the study: Business Economics—39 students (31.2%), Faculty of Tourism and Hospitality Management—30 students (24%), Faculty of Informatics and Computing—17 students (13.6%), Faculty of Physical Education and Sports Management—1 student (0.8%), Anglistics—29 students (23.2%), and Pharmacy—9 students (7.2%). The surveyed students were from various years of undergraduate studies and master's programs. The sample included first-year, second-year, third-year and fourth-year undergraduate students, as well as 3 (2.4%) master's students. All respondents studied two foreign languages. All students at Singidunum University learn English, while a second foreign language was chosen as an option (Spanish, French, Russian, German, Italian, Greek, or Chinese).

The teaching staff who participated in the study consisted of foreign language teachers with diverse academic qualifications and work experience. Among the 25 surveyed teachers, 21 (84%) were female and 4 (16%) were male. Most teachers taught English at the university—14 (56%), while teachers of other foreign languages (Spanish, French, Russian, German, Italian, Greek, and Chinese) comprised 11 (44%) respondents. The teachers had varying academic titles and lengths of teaching experience. The work experience of the respondents varied: up to 5 years—2 (8%) respondents, 5 to 10 years—2 respondents (8%), 10 to 20 years—13 respondents (52%), and over 20 years—8 (32%).

If we think about communication difficulties in hybrid or online teaching models, the first thing that usually comes to our mind are technical problems, ranging from power outages and internet disruptions to issues with devices used to participate in classes. We examined the opinions of both teachers and students and posed



the following questions. The results are presented in Table 1. The question for both students and teachers was: *During foreign language classes in a hybrid model, do technical problems disrupt learning (quality of internet traffic, computer issues, power outages)?*

By analyzing the results, we can conclude that, for many teachers, technical problems affect the flow of the teaching process, while only about twenty percent of students believe that technical problems disrupt the learning process. We believe this outcome is due to the high level of digital competence among students, as they quickly resolve technical issues (e.g., using a mobile phone to connect to the class during a home internet outage). On the other hand, teachers lose significantly more time restoring an uninterrupted work process after an internet network failure in the classroom, as they rely on institutional devices and internet connections.

Table 2 presents the forms of class attendance, showing that only a quarter of the surveyed students always attend classes in the classroom. Based on this, we can see that most students often or occasionally attend online classes, and their responses are based on their personal experiences of participating in online foreign language classes.

2.1. CLASS INTERACTION IN A HYBRID TEACHING MODEL

2.1.1. Differences in teacher-student interaction in the classroom and online

We assume that the teacher pays more attention to students in the classroom than to those online. We wanted to check this assumption by asking two similar questions, one for students, and one for teachers. The question for students was: 'The teacher pays more attention to students attending classes in person and less to those who are online, which affects my motivation to participate in class.' The question for teachers was: 'One of the biggest challenges during the implementation of a hybrid class model is equally perceiving the feedback of students present in the classroom and those on the online platform.' The obtained results are presented in Table 3.

Table 1. The influence of technical problems on disrupting the process of foreign language learning and teaching in the hybrid teaching model

Perspective	Students	Teachers
Strongly Disagree	31,2%	4%
Mostly Disagree	28,8%	12%
No Opinion/Neither Agree nor Disagree	18,4%	8%
Mostly Agree	13,6%	14%
Strongly Agree	8%	52%

Table 2. Form of attendance in foreign language classes

	Students
Always in the classroom	25,6%
Always online	18,4%
Equally in the classroom and online	16%
More often in the classroom	14,4%
More often online	25,6%

Table 3. The teacher's focus on students in the classroom and online

Perspective	Students	Teachers
Strongly Disagree	44,8%	4%
Mostly Disagree	20,8%	8%
No Opinion/Neither Agree nor Disagree	12,8%	8%
Mostly Agree	15,2%	40%
Strongly Agree	6,4%	40%



What can be concluded based on the given results is that the majority of students believe that the teacher has more intensive interaction with students who are physically present in the classroom, which confirms our hypothesis. Of course, it should be emphasized that the surveyed students attend classes with several foreign language teachers, depending on their year of study and the study program. Therefore, we can conclude that their impressions are the result of comparing classes conducted by different teachers, making their conclusion relevant since it is not based on observing the lessons of just one foreign language teacher.

It is evident from the results from the teachers' questionnaire that the interaction between teachers and students in the classroom and students present online is uneven. More precisely, 80% of the surveyed teachers completely or mostly agree that they do not perceive feedback equally from students in the classroom and those online. Therefore, the interaction is more intensive with the group that the teacher focuses on at a given moment.

2.2. DIFFERENCES IN INTERACTION BETWEEN STUDENTS IN THE CLASSROOM AND THOSE PARTICIPATING ONLINE

There is almost no interaction between students in the classroom and online; it is sporadic and rare. This may have some influence on the quality of language teaching as there is no communication among all participants in the teaching process. Students attending classes online cannot see the students participating in the classroom.

To check the attitude of language teachers, we asked the following question: 'During hybrid foreign language

classes, interaction between students who are present in the classroom and those online is very difficult or impossible to achieve.' The obtained results are presented in Table 4. The attitudes of students were analyzed through questions regarding disruption in non-verbal communication, which we will discuss further below.

2.3. DISRUPTION DUE TO THE LACK OF NON-VERBAL COMMUNICATION

Our hypothesis is that disruption and hindered communication develop during hybrid teaching for several reasons. The teachers' attention is scattered as they are simultaneously focused on both groups of students. Due to this fact, they are unable to assess the engagement and participation of each student. This is further exacerbated by the fact that most students participating online do not turn on their cameras or microphones, which results in a complete absence of non-verbal communication.

We analyzed the extent to which it is important for students to see one another during class, to observe facial expressions and gestures, and we obtained the following results. The statement we made in the survey was: 'When I attend foreign language classes, I miss seeing my colleagues participating in class from the classroom because observing the facial expressions and gestures of the speaker is very important to me in oral communication.' The statement teachers were asked to respond to was: 'Students attending online foreign language classes without turning on their cameras reduce the level of interaction because their facial expressions and gestures are missing.' The obtained results are presented in Table 5.

Table 4. Interaction between students in the classroom and those participating online from the teachers' perspective

Perspective	Teachers
Strongly Disagree	12%
Mostly Disagree	12%
No Opinion/Neither Agree nor Disagree	0%
Mostly Agree	64%
Strongly Agree	12%

Table 5. The importance of using cameras during foreign language classes in the hybrid model

Perspective	Students	Teachers
Strongly Disagree	25,6%	4%
Mostly Disagree	20,8%	8%
No Opinion/Neither Agree nor Disagree	32%	0%
Mostly Agree	15,2%	44%
Strongly Agree	6,4%	44%

**Table 6.** The influence of using a camera on students' concentration and comfort during class

	Students
Strongly Disagree	11,2%
Mostly Disagree	11,2%
No Opinion/Neither Agree nor Disagree	20%
Mostly Agree	26,4%
Strongly Agree	31,2%

What can be concluded from these results is that the absolute majority of teachers believe that the use of cameras by online students is essential for a high level of interaction, whereas only 20.6% of students consider the use of cameras significant for the level of interaction. It is interesting to note that the majority of students do not want to turn on their cameras during foreign language classes, and some even refuse to use a microphone. We assumed that the reason for this is the feeling of discomfort when the camera is on. Logically, when a student feels uncomfortable, focusing on the lecture becomes more demanding. From the survey results, that are presented in Table 6, we can see that more than half of the students confirmed that they feel uncomfortable when their camera is on during class.

3. CONCLUSION

Based on the study on disruption in the hybrid model of foreign language teaching at Singidunum University, several key conclusions can be drawn. First, it has been confirmed that the hybrid model brings numerous advantages, but also significant challenges in communication between teachers and students. The lack of eye contact, hindered interaction, and reduced non-verbal communication are some of the issues affecting the quality of teaching. Students attending classes online often feel excluded compared to their peers in the classroom, while teachers encounter difficulties in focusing their attention and maintaining the dynamics of the lecture. From a technical standpoint, problems were identified as a significant factor that can disrupt the teaching process. Although students demonstrate high digital literacy and overcome these issues more easily, teachers perceive them as a serious challenge, particularly due to the use of institutional equipment and internet connectivity. The analysis of interaction revealed that teachers are more focused on students attending classes in the classroom, which can potentially negatively impact those joining online. This situation may lead to decreased engagement of students who follow classes online.

Finally, the research results indicate the need to improve the hybrid teaching model through enhanced technical infrastructure, better quality interaction, and greater integration of digital tools. Teachers should develop techniques that ensure equal teaching quality for all students, regardless of their mode of attendance. It is necessary to work on the development of this teaching model to ensure effective and high-quality communication among all participants in the educational process.

4. ACKNOWLEDGMENTS

We would like to dedicate this paper to our dear late colleague, Jovan Travica.

REFERENCES

- [1] M. Krstinic, M. Pauković, "Hibridna nastava- izazov i mogućnost novog normalnog," *Obrazovanje za poduzetništvo*, pp. 130-139, 2021. , <https://doi.org/10.38190/ope.11.2.7>
- [2] V. Bošković Marković, M. Veljković Michos, D. Lunić, K. Nasradin and M. Pupavac, "Disadvantages of foreign language teaching at the University level," *Sinteza 2022*, pp. 341- 346, 2022., <http://dx.doi.org/10.15308/Sinteza-2022-341-346>
- [3] K. C. O. Iermachkova, "Role of Digital Technologies in the Foreign Language Classroom," *Digitalization of Education: History, Trends and Prospects*, pp. 538-543, 2020. , <https://doi.org/10.2991/assehr.k.200509.098>
- [4] D. Stankovic and L. Stoimenov, "Hibridni model nastave uz primenu najnovijih informaciono-komunikacionih tehnologija," *18. Telekomunikacioni forum TELFOR*, pp. 23-25, 2010. , http://2010.telfor.rs/files/radovi/TELFOR2010_10_22.pdf
- [5] "Metodika i komunikacija e- obrazovanja," CAR-NET, Hrvatska akademska i istraživačka mreža., <http://edupoint.carnet.hr/referalni/obrazovni/mkod/metodika/hibridno.html>



- [6] B. F. Klimova, "Impact of ICT on foreign language learning," *A WERProcedia Information Technology*, pp. 180-185, 2012. , <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=bf743053a91089f1d6467a6cb42ace2f240182ae>
- [7] A. J. Guerrero-Quíñonez, M. Bedoya- Flores, E. Mosquera-Quíñonez, E. D. Ango-Ramos and R. M. Lara-Tambaco, "Hybrid education: Current challenges," *Ibero- American Journal of Education & Society Research*, vol. 3, no. 1, pp. 276-279, 2023.
- [8] E. R. Eslit, "Teacher-Student Interaction in Higher Education Institutions: Hybrid Language Learning in the Post- Pandemic Landscape," *Preprint. Arts and Humanities*, vol. 1, 2023. , <https://doi.org/10.20944/preprints202310>



AI-GENERATED TEXTS IN A FOREIGN LANGUAGE CLASS: PROS AND CONS

Vladimir Fedorov*,
[0009-0005-2235-8876]

Miloš Pupavac,
[0000-0002-6179-3900]

Maja Rončević
[0000-0002-5743-9124]

Singidunum University,
Belgrade, Serbia

Abstract:

The aim of this paper is to investigate whether artificial intelligence tools can be used in text generation, as a basic teaching tool, in foreign language classes. The request listed words and phrases from which a text in the form of a monologue and a task with statements related to the text were to be generated. The texts and tasks were generated using ChatGPT and Gemini in English, Russian and Serbian. Authenticity, grammatical correctness, vocabulary, spelling, tasks and their usability in teaching were analyzed. The conclusions were reached that artificial intelligence tools can be used to generate text for a foreign language class, but they must be reviewed and corrected if errors in grammar and spelling appear. The generated texts and tasks are very useful because they provide ideas and generate teaching materials that can motivate both the teacher and those learning a foreign language.

Keywords:

AI, Text, Foreign Languages, Generating Text, Input.

INTRODUCTION

Using AI in foreign language teaching has long been the subject of both fervent debate and eager anticipation—well before AI tools reached their current level of sophistication, making it possible to incorporate them into day-to-day teaching practice. Today, numerous AI-based resources are available online to support teachers in their work, ranging from tools that help check awkward-sounding sentences, such as <https://ludwig.guru/>, to comprehensive toolkits for building teaching materials from scratch, like <https://twee.com/>.

However, despite the wide range of AI applications already present in education, doubts remain—not only about its usefulness but, more importantly, about its reliability in terms of linguistic, cultural, and even logical or common-sense adequacy and accuracy.

On the one hand, there are many positive voices: “AI offers a plethora of novel educational insights, methodologies, and resources, empowering educators with innovative tools to enhance language learning outcomes.” [1]

Correspondence:

Vladimir Fedorov

e-mail:

vfedorov@singidunum.ac.rs





Some even view AI as an essential or permanent feature of the future classroom: “The classroom of the future will cleverly combine the advantages of digital learning with proven, computer-free methods, content, and tasks for face-to-face teaching, which will remain indispensable and highly significant for successful learning.” [2] At the same time, others express voicing “a significant concern” caused by “the potential over-reliance on ChatGPT.” [3]

The goal of this paper is to explore one of the most interesting and potentially beneficial ways AI can be used in the classroom. Extensive, high-quality input is essential for mastering a language, particularly because key aspects like vocabulary acquisition and natural usage depend on it: “A good part of vocabulary acquisition has to be incidental. Incidental learning is facilitated through exposure to language input, in the form of extensive reading, for example.” [4]

As a result it seems a very obvious step to start using AI not only for creating tasks, questions, lesson plans etc., but also to make it produce high-quality reading and listening input. Never before did there exist a possibility of generating material which could be so perfectly tailored to the requirements a teacher might have both in the sense of its relevance to the student and its linguistic content.

We would like to focus on reading input not only because it is essential for mastering a language, but also because of the role it plays in developing general cognitive adeptness. The latter is eloquently illustrated in a detailed explanation of the reading process from a comprehensive guide by J. Willis, M.D.: “These are the parts of the reading process when the brain links the abstract orthographic representations it decodes with its system of phonological codes. This is when patterning begins to take the decoded words and process them into comprehensible categories, and when words and phrases are associated with meanings in the process of developing fluent reading. Simultaneously, *word vocabulary is increasing* and strategies are available to *facilitate vocabulary-building skills*. Ultimately, the patterning of phonological coding, enriched by greater vocabulary, combines with the *increased fluency* to reach the later reading stages of comprehension of increasingly complex text. [*Emphasis in all cases added*].” [5]

Obviously, reading plays a crucial role in developing several vital competencies for a foreign language learner, with vocabulary being one of the most significant: “One of the primary skills a second language learner needs to develop is the knowledge of vocabulary, a skill emphasized by many researchers as a crucial feature of

language learning,” while “many linguistic investigations and research have pointed out that there is a close connection between reading proficiency and vocabulary knowledge.” [6] Moreover, “the text... simultaneously represents both a tool and an object of teaching: a tool—insofar as it introduces new and unknown (mainly grammatical and lexical) elements,” [7] thereby offering a valuable opportunity to address multiple aspects of language competence. Therefore, beginning with such a fundamental activity as reading appears to be a solid starting point for exploring AI’s potential in generating input for the foreign language classroom.

In selecting (creating) material it is important to keep in mind that “the most significant factors in the reading selection process are related to the students: the students’ level, interests, needs, and background knowledge. Other factors are related to the text itself: content, relevance and authenticity.” [8] It seems we have every reason to expect artificial intelligence to meet both sets of criteria. In fact, no textbook or other traditional source can come even remotely close to the level of customization that AI is—at least potentially—capable of providing.

A long-standing discussion worth mentioning concerns whether to use authentic, adapted, or synthetic texts in the classroom: “Authentic texts were not created for the purpose of fulfilling some didactic goal, while synthetic texts are compiled by textbook authors and foreign language teachers, most often with the intention of illustrating a linguistic phenomenon.” [7]

In our view, there is a strong case for using adapted texts, particularly in the early stages of language learning. The advantages of this approach can be summarized as follows: “Adapting materials can make them [texts] accessible, interesting, and informative. – Arranging materials around one theme allows the learner to build up background knowledge of the content... – If the text is poorly presented, it can always be improved... – They allow teachers to evaluate their students.” [8]

The type of reading we are aiming for is what is commonly referred to as **intensive**—or, less commonly, **exploratory**—reading. “Exploratory reading is one of the main types of reading used in learning... a foreign language. Its goal is to maximize the full perception and understanding of the text being read. Reading takes place at a slower pace and is accompanied by rereading of individual text fragments. This type of reading involves... a whole system of exercises and tasks related to different levels of understanding of the text after reading, with progression to oral and written speech.” [9]



To summarize, the goals we aim to achieve are to ensure that the generated text supports the development of various linguistic competencies, such as listening comprehension, reading comprehension, and speaking. For the purposes of our analysis, we created a text in the form of a monologue tailored to the beginner level of foreign language learning, focusing on an everyday topic—namely, daily activity schedules. According to Ginić and Ajdžanović, the length of the text must adhere to specific guidelines, which is why the length of the monologue is set at 120–150 words. [10]

2. ANALYSIS

To create texts and tasks using artificial intelligence tools, we employed two platforms: ChatGPT by OpenAI and Gemini by Google. We provided the same input in Serbian, specifying the elements that the text should include, along with specific phrases. For instance, we emphasized that, when it comes to names and meals, the text should reflect foods and names typical of the country where the language is spoken (in our case, Serbia, the Russian Federation, and Great Britain). Our task was:

“Can you write a monologue with a typical name for Serbia (Russian Federation/Great Britain) on the topic “My Day” in Serbian (Russian/English)? The text should be at A2 level and contain 150 words. The names and food should be typical for Serbia (Russian Federation/Great Britain). The speaker is about 20 years old. The text should include words such as: training, showering, television, waking up, getting up, making, preparing, cooking, breakfast, drinking coffee/tea, getting ready for school, lectures, break, café, socializing, returning home, rest, washing up, dinner, going to bed, shopping, washing dishes, pet, brushing teeth, getting dressed, family.”

After generating the text, we submitted the same input, adding:

“Can you create a task that includes statements to be answered as true/false, which would check whether the student understood the text well?”

When assessing the generated texts, we focused on the following considerations:

- A) Authenticity – We analyze whether the sentences sound natural and reflect the language as it is used in communication. Our expectation is that some sentences may not sound natural to a native speaker.

- B) Logical Flow of Activities – Given the illogical sequence in our input, the listed activities may not follow a typical daily schedule. We do not expect all activities to be arranged in a logical order.
- C) Grammar and Spelling – We anticipate grammatical and spelling errors, particularly in texts in Serbian and Russian.
- D) Translation of Phrases and Expressions – Since the instructions are in Serbian, we expect literal translations of phrases that may not be used naturally in the target language.
- E) Cultural Elements – We expect that the character’s name and the food mentioned in the text will be authentic to the culture in which the language is spoken.
- F) Task – We expect the generated statements related to the text will be clear and motivating for the adoption of new language material. However, we foresee some grammatical or semantic issues in these statements.

2.1. REVIEW OF AI-GENERATED MATERIAL IN SERBIA

When it comes to text length, we initially set the monologue to contain 150 words. However, the actual word count varied: ChatGPT generated a text with 115 words, while Gemini’s text contained 124 words. Regarding authenticity, the texts were notably different. The text generated by Gemini contained sentences that aligned well with native speaker expectations, whereas the text generated by ChatGPT exhibited an unexpected word order in sentences containing reflexive verbs.

Although our request lacked logical sequencing in the listed activities, both texts presented them in a coherent and logical order. This confirms our initial assumption.

In terms of grammar and spelling, Gemini’s text contained no grammatical errors but included two redundant commas, as seen in the following examples: “По повратку кући, одморим се мало уз телевизију” and “Пре спавања, перем зубе и читам књигу.” In contrast, ChatGPT’s text contained grammatical errors related to word order in reflexive verb constructions. For instance, in the sentence “Свако јутро будим се у 7 сати,” the reflexive “се” should precede the verb “будим.” Similarly, in “После факултета, враћам се кући [...]” the reflexive pronoun should also be placed before the verb: “После факултета се враћам се кући [...].”



Regarding phrases and expressions, no significant errors were detected, as the input was in Serbian.

The cultural aspect of the generated texts is particularly interesting. Both AI models selected common Serbian names: Gemini's text featured Miloš, while ChatGPT's featured Marko, both of whom are 20 years old. Additionally, the breakfast choices reflected Serbian culinary traditions. ChatGPT included eggs and bread with cheese, while Gemini mentioned *proja* with cheese, a dish with deep roots in Serbian cuisine. Furthermore, Gemini's text described Miloš starting his morning with strong traditional ("homemade") coffee, a widespread habit among Serbs.

Both tools generated statements that can effectively engage students in discussions. ChatGPT produced ten statements, whereas Gemini generated eight, both concluding with correct answers. Notably, Gemini incorporated traditional dishes such as *gibanica* and beans in its statements, providing an opportunity to expand students' lexical knowledge and cultural awareness.

In conclusion, AI tools like ChatGPT and Gemini can be valuable assets for teachers in lesson preparation. However, it is essential to review and refine the generated material before using it in class to ensure linguistic accuracy and cultural appropriateness.

2.2. REVIEW OF AI-GENERATED MATERIAL IN RUSSIAN

Both ChatGPT and Gemini produced texts which maintained a natural flow and were generally accurate in terms of real-life details, with a few notable exceptions. They chose the name Ivan which is indeed one of the most common male names in Russian. However, several issues are worth mentioning.

Probably the most striking was the fact that both AI models initially generated texts in Serbian despite an explicit request to use Russian. The same issue occurred in the subsequent task. An additional request was needed for them to change the language.

ChatGPT generated a text of 132 words. It made one major semantic error: while telling about his daily routines, Ivan said "варю суп и пиццу," literally "[I] boil soup and pizza." While you can "boil soup" in Russian (it is a standard phrase), you certainly cannot do that with pizza. The mistake was repeated in the True/False task after the text: "Вечером Иван варит суп или пиццу" (lit. "In the evening Ivan boils soup and pizza"). Another serious mistake was grammatical and appeared in the task instructions: "отметить каждое утверждение как

ПРАВИЛЬНО или НЕПРАВИЛЬНО в соответствии с информацией в тексте," lit. "mark every statement as true or false according to the information in the text." The so called short forms of adjectives were used where the full ones would be expected ("отметить каждое утверждение как ПРАВИЛЬНОЕ или НЕПРАВИЛЬНОЕ"). Moreover, it is much more common to use the words ВЕРНО(Е) and НЕВЕРНО(Е) respectively, even though ChatGPT's choice is not strictly incorrect. Interestingly, Gemini used the more natural wording.

Both ChatGPT and Gemini also used the word "питомец" ("pet"). While neutral and fairly common, its usage is typically limited to specific contexts such as pet shops or veterinary clinics. For instance, it appears in "карта питомца" (a pet's record at a vet hospital), or when an adult addresses a child he or she does not know very well ("А у тебя есть какой-нибудь домашний питомец?" — "Do you have a pet?"). In other words, питомец always implies a certain distance (which by no means necessitates formality, however) between the addresser and the addressee.

Gemini generated a piece of 140 words. The model had its own issues. At the beginning, Ivan says "учусь в Москве" ("[I] study in Moscow"), which sounds somewhat unusual since it is more common to specify the educational institution rather than the city unless there is a particular reason to do so. Additionally, "блины" (Russian pancakes) are highly uncommon as a weekday breakfast. The phrase "готовлю ужин для семьи" ("[I] cook dinner for [my] family") also sounds unnatural: "готовлю ужин для всей семьи" ("for the whole family") or "готовлю ужин для своей семьи" ("for my family") would be more appropriate. Interestingly, Gemini corrected itself in the task section, suggesting the more natural phrase "Иван готовит ужин для своей семьи."

Furthermore, while Masha (a diminutive of Maria) is a common female name in Russia, it is almost never used for cats or dogs. Like ChatGPT, Gemini made a grave grammatical error in the task: "У Ивана есть собака в качестве домашнего питомца" (lit. "Ivan has a dog as a pet") instead of "У Ивана есть домашний питомец — собака".

In conclusion, both ChatGPT and Gemini performed well overall and can certainly be useful for generating content at this level. However, careful double-checking is necessary, even for grammar, let alone culturally specific details. In fairness to AI, the issue with питомец, is quite subtle as this nuance is rarely, if ever, explicitly articulated in Russian dictionaries or textbooks.



2.3. REVIEW OF AI-GENERATED MATERIAL IN ENGLISH

The AI-generated texts in English, created by ChatGPT and Gemini, meet the basic needs of beginner to pre-intermediate ESL learners. Both texts are clear, grammatically correct, and suitable for classroom use. However, there are some small differences in style, tone, and how useful they are for teaching.

Both texts followed a good structure. ChatGPT's text had 159 words, while Gemini's was a bit longer with 172 words. Each one described a typical daily routine in a logical order. The sentence structure was simple, using the present simple tense, which is appropriate for this level of learners.

Gemini's text sounds more relaxed and friendly. It starts with "Hi everyone!" and uses expressions like "It's delicious!", which make it more like spoken English. ChatGPT's version is more formal and direct, starting with "My name is James, and I am 20 years old." This can be helpful when learners need clear and simple language.

Both texts include British cultural elements. ChatGPT mentions "fish and chips" and "eggs and toast," while Gemini adds more traditional British foods like baked beans, scones with clotted cream, and English breakfast tea. Gemini also mentions "shepherd's pie," which can be used to teach new words and talk about culture in class.

Grammatically, both texts are mostly correct. However, Gemini's use of "Marmite" as the name of a cat could confuse some learners who don't know it's also the name of a British food spread. ChatGPT used a more typical name — Max — which is easier for students to understand. Also, Gemini's phrase "I usually cook something simple" is a nice and useful sentence that learners can copy when talking about their own meals.

The true/false questions in both tasks match the texts well and are easy to follow. Gemini's questions are a little shorter, but both sets are clear. The false statements, like saying Oliver wakes up at 6 a.m. or has a dog, are realistic and help students think about the details in the text.

In conclusion, both ChatGPT and Gemini created texts that are useful for English language teaching. Gemini's version is more fun and casual, while ChatGPT's is more structured and formal. Depending on the goals of the lesson, teachers can choose the one that suits their needs best. However, it is always a good idea for teachers to check and adjust the texts before using them in class, especially to make sure the words and cultural references are clear to students.

3. CITATIONS

To summarize, some of our expectations were confirmed, while others were not. Regarding authenticity, although a few elements were somewhat unnatural, the majority were well-formed. Logical consistency was excellent, despite the fact that the listed activities appeared in an incorrect order. Minor grammatical issues were present, as well as a punctuation problem in Serbian. However, spelling was entirely accurate. There were no instances of literal translation, even though the request was made in Serbian. That said, the issue of generating the entire content in Serbian first, rather than in Russian, persisted. Cultural elements were incorporated in all cases, though not always with full accuracy. The subsequent task was well-structured and appropriate for use, though, as expected, not entirely free from errors.

One particularly interesting observation is that AI tools provide teachers with ideas to reflect on when preparing for class. In our case, they suggested culturally specific and popular food items—ideas that were both relevant and creative, even if not immediately obvious. Overall, Gemini performed better, both in terms of producing fewer errors and in structuring tasks more effectively. For instance, it introduced a wider range of vocabulary at the appropriate level, closely related to the words in the text, yet without exceeding the requested difficulty (e.g., pairing "cat" with "dog"). Additionally, it provided a greater number of words across all three languages.

Our conclusion and recommendation are that AI tools can be a valuable resource for generating input in foreign language classes. They produce generally high-quality content and, more importantly, generate useful ideas. However, it remains the teacher's essential responsibility to thoroughly review the output for semantic, grammatical, and cultural accuracy.

REFERENCES

- [1] A. Yang, "Challenges and opportunities for foreign language teachers in the era of artificial intelligence," *International Journal of Education and Humanities*, vol. 4, no. 1, pp. 39-50, 2024.
- [2] T. Schmidt and T. Strasser, "Artificial intelligence in foreign language learning and teaching: a CALL for intelligent practice," *Anglistik: International Journal of English Studies*, vol. 33, no. 1, pp. 165-184, 2022.



- [3] F. Karataş, F. Y. Abedi, F. Ozek Gunyel, D. Karadeniz and Y. Kuzgun, "Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners," *Education and Information Technologies*, vol. 29, p. 19343–19366, 2024.
- [4] S. Thornbury, "How to teach vocabulary", Pearson Education India, 2006.
- [5] J. Willis, "Teaching the brain to read: Strategies for improving fluency, vocabulary, and comprehension", Alexandria, VA, USA: Ascd, 2008.
- [6] V. Gavranović, "Second Language Vocabulary Acquisition Enhanced by the Use of Technology," in doi:10.15308/Sinteza-2019-621-626, Belgrade, 2020.
- [7] O. Durbaba, „Reafirmacija autentičnog teksta kao novog (starog?) bazičnog elementa u nastavi stranih jezika," *INOVACIJE u nastavi*, т. XIX, бр. 3, pp. 16-28, 2006.
- [8] I. J. Arias, "Selecting reading materials wisely," *Letras*, no. 41, pp. 131-151, 2007.
- [9] Л. С. Крючкова and Н. В. Мощинская, "Практическая методика обучения русскому языку как иностранному", Москва: Флинта, 2009.
- [10] J. Т. Гинић и Н. Д. Ајџановић, „Методологија организације лингводидактичког материјала у уџбенику руског језика "Ни пуха ни пера!", Филолог, бр. 23, pp. 132-149, 2021.



INFORMATION TECHNOLOGY IN SPORTS SESSION



THE GROWING ROLE OF NUTRITION APPLICATIONS IN HEALTH AND WELLNESS - THE DEVELOPMENT OF CISHRANA, SERBIA'S FIRST NUTRITION APPLICATION

Marija Anđelković*,
[0009-0009-7220-2617]

Nenad Dikić
[0000-0001-5245-2891]

Singidunum University,
Belgrade, Serbia

Abstract:

With increasing awareness of the connection between diet and health, nutrition applications have become essential to promoting balanced eating habits. These apps provide real-time dietary tracking, calorie counting, and nutrient analysis, helping users make informed food choices. The rising prevalence of obesity and diet-related diseases has driven the need for digital solutions that empower individuals to monitor their nutritional intake and achieve health goals. This paper explores the functionalities and benefits of nutrition applications, their role in weight management and overall wellness, and their challenges. Additionally, we present the development of CISHrana, Serbia's first nutrition application, designed to offer precise dietary guidance based on a localised food database. The CISHrana app results from a collaboration between the Center for Sports Nutrition and Supplementation (CIS) and the Synchronized Swimming Federation of Serbia. This pioneering initiative provides Serbian users with a scientifically backed dietary tracking tool. The paper highlights the app's features, development process, and potential impact on public health and professional nutrition practice.

Keywords:

Nutrition Apps, Health Technology, Digital Health, Dietary Tracking, Public Health Innovation.

INTRODUCTION

Maintaining a balanced diet has become increasingly challenging due to fast-paced lifestyles, sedentary habits, urbanisation, and the widespread consumption of processed foods. Highly processed foods, often rich in sugars, unhealthy fats, and artificial additives, have contributed to rising rates of obesity and diet-related diseases globally [1] [2]. In response to these challenges, nutrition applications have emerged as essential tools that help users track food intake, manage calories, and plan their diets. These digital platforms enable users to monitor their eating habits, analyse nutrient intake, and receive personalised recommendations to achieve specific health goals [2].

With chronic conditions such as obesity, diabetes, and cardiovascular diseases on the rise, personalised nutrition guidance is more critical than ever. Studies indicate that nutrition apps improve adherence to dietary recommendations by providing real-time feedback and educational resources [3] [4] [5].

Correspondence:

Marija Andjelković

e-mail:

mandjelkovic@singidunum.ac.rs





Advanced nutrition apps leverage machine learning algorithms, artificial intelligence, and comprehensive food databases to offer individualised meal plans, making them more effective than traditional dietary counselling alone [6].

Additionally, modern digital nutrition tools increasingly integrate biometric data from wearable devices, allowing users to track physiological responses to their dietary choices in real time [7].

Several global nutrition apps, including MyFitnessPal, Yazio, and Cronometer, have effectively promoted weight management and healthier eating habits [3] [8] [9]. However, many of these applications rely on Western food databases, limiting their suitability for non-Western populations due to differences in dietary patterns and available food items [9]. CISHrana was developed to address this gap, offering the first Serbian nutrition application to accommodate local nutritional habits and food availability [10]. By integrating a locally sourced food database, CISHrana enables accurate meal tracking and planning aligned with Serbian users' dietary needs.

This paper explores the role of nutrition applications in modern health and wellness, their influence on dietary habits, and the challenges they face. Additionally, it introduces the development of CISHrana, detailing its functionalities, objectives, and anticipated contributions to public health.

1.1. DEVELOPMENT OF THE CISHRANA APP

Recognising the demand for localised nutrition solutions, the Center for Sports Nutrition and Supplementation (CIS - <https://cis.edu.rs/>) partnered with the Synchronized Swimming Federation of Serbia (<https://www.synchro-serbia.org.rs/index.php>) developed CISHrana. This initiative was part of the FINA OASP programme, which promotes athlete development through nutrition education. CISHrana was explicitly designed to meet the dietary needs of Serbian users by incorporating local food databases, including EuroFIR/FoodEXplorer. The app lets users track caloric intake, analyse meal composition, and receive personalised dietary recommendations.

Despite the development of mobile applications aimed at monitoring nutritional intake, most tools available globally are designed based on foreign food databases, lack culturally relevant items, and are not adapted to the specific dietary habits, items and portion sizes typical of the Serbian population. In response to this gap, we developed the first Serbian mobile application to serve

as a user-friendly, scientifically grounded nutrition diary. The app is available on Android and iOS platforms. It includes an extensive database of Serbian food items with localised portion measurements and nutritional values derived from verified national and international sources. It enables users to log daily meals, monitor macronutrient and micronutrient intake, and gain personalised insights into their dietary patterns.

CISHrana integrates national food data, ensuring accurate nutritional assessments. It is designed for athletes and the general population and features an intuitive interface that simplifies dietary tracking. The application was developed with the assistance of Alchemy Engine experts (<https://www.alchemyengine.io/>), ensuring a seamless user experience. It is available on Android and iOS platforms, with planned updates to enhance functionalities based on user feedback and ongoing nutritional research.

The app includes several key features, including:

- A meal-tracking system that calculates daily macronutrient and micronutrient intake;
- A visual representation of nutritional summaries;
- A detailed food database with Serbian-specific items.



1.2. EXAMPLE OF THE APP INTERFACE

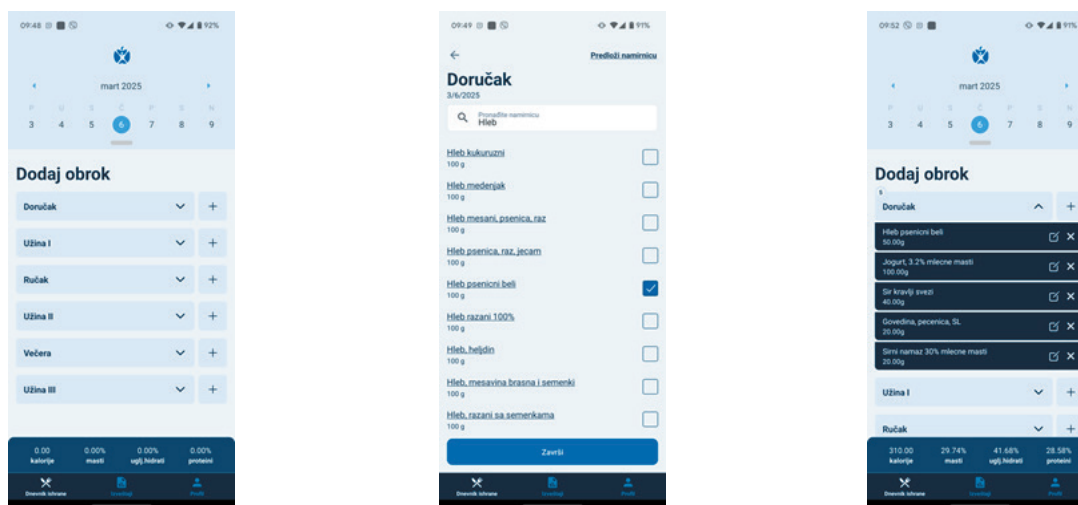


Figure 1. CISHrana - Meal intake tracking interface

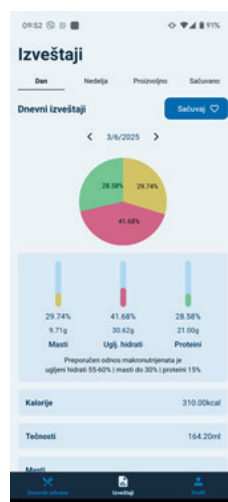


Figure 2. CISHrana - Summary of daily caloric and nutrient intake

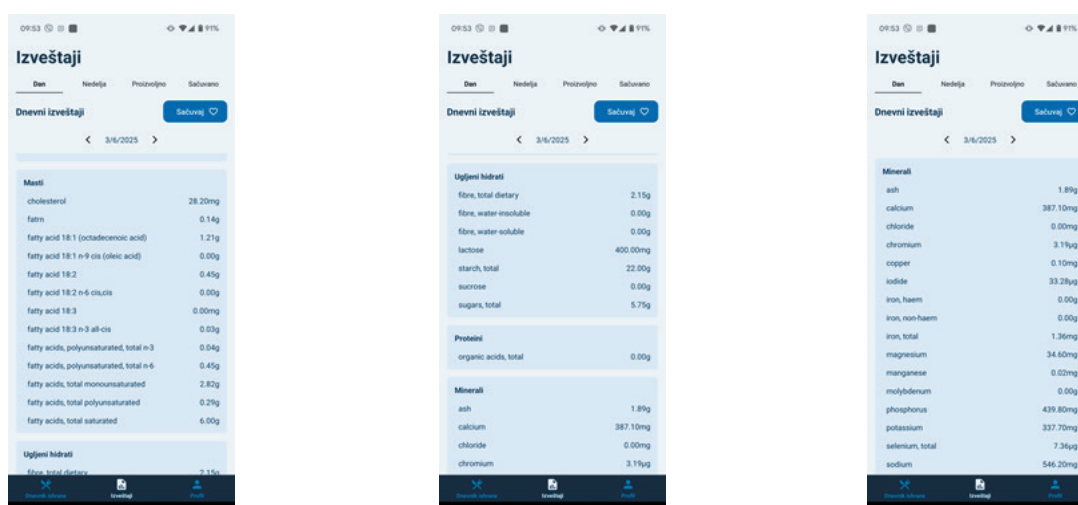


Figure 3. CISHrana - Detailed nutritional report



The app interface presented in Figures 1, 2, and 3 shows the app's usability, accuracy, and potential applications in clinical practice, sports nutrition, public health education, and personal health monitoring. The user can choose from thousands of different nutrition items, adjust the dosage, and add them to one of five meals (breakfast, lunch, dinner, and snacks). In Figures 2 and 3, the daily report with represented macro and micronutrient values is easy for users to read and determine how many calories, fluids, vitamins, and minerals are in every meal.

This paper aims to present the application's structure, development process, and validation methodology, with a focus on its usability, accuracy, and potential applications in clinical practice, sports nutrition, public health education, and personal health monitoring. We discuss the relevance of integrating culturally tailored digital health tools into everyday medical and lifestyle practice and the broader implications for digital transformation in nutrition science.

2. DISCUSSION

Modern nutrition applications are evolving into sophisticated tools that facilitate healthier eating habits through technology. The primary function of most nutrition apps is to enable users to log their daily food intake and receive feedback on calorie and nutrient consumption. However, it is well-established that food items vary significantly in macronutrient and micronutrient content across different regions. Regional differences in sugar content, fortified nutrients, and preservatives can substantially affect the accuracy of global food databases. In response, CIShrana utilises the EuroFIR food database, incorporating data from local food items to ensure the precision of dietary calculations. This app is further supported by information from *Calories in Everyday Life*, a comprehensive book by Dr. Nedeljko Jokić, which presents tables of local food ingredients, further enhancing its regional relevance.

Scientific research supports the notion that individuals who track their food intake are more likely to make healthier choices and maintain balanced diets [3]. By incorporating a locally adapted food database, CIShrana offers a user-friendly approach that allows individuals to search for familiar, locally sourced food items in the local language. This makes it not only accessible but practical for Serbian users who are accustomed to their regional dietary habits.

The development of this application represents the culmination of over a decade of dedicated work in sports nutrition, driven by the Center for Sports Nutrition and Supplementation (CIS). Although the concept of creating a calorie and nutrient tracking app had been in the works for several years, it was not until 2023, through a collaboration with the Synchronised Swimming Federation of Serbia and the support of the FINA OASP programme, that the initiative came to fruition. This international programme, which supports the development of aquatic sports, provided both the financial and organizational resources needed to turn this vision into reality. The app was initially conceived as an educational tool for athletes, particularly those in aesthetic sports such as synchronised swimming, where nutrition and body image are essential. It was designed to support athletes in managing their nutritional intake more effectively. Through nutrition workshops and educational lectures provided by CIS experts, it became apparent that an intuitive, mobile platform was the most practical solution for translating complex concepts such as energy balance and portion control into athletes' daily routines. Unlike generic nutrition applications that serve a broad global audience, CIShrana is explicitly tailored to the Serbian population. This focus ensures that users receive dietary assessments that reflect local food ingredients, portion sizes, and standard nutritional practices. While many nutrition apps incorporate barcode scanners for packaged food items, they often rely on international databases, which may not accurately reflect the nutritional composition of locally produced foods. As part of its ongoing development, CIShrana aims to integrate a barcode scanning feature specifically for Serbian food products, further enhancing its accuracy and usability.

The future of nutrition tracking applications lies in their ability to integrate with wearable fitness devices, such as Fitbit, Apple Watch, and Google Fit. These integrations will enable a comprehensive view of a user's health by linking dietary intake with physical activity levels, creating a more holistic approach to health management. CIShrana is designed with a modular framework that allows for future updates, including such integrations, ensuring the app remains at the forefront of digital health solutions.

In addition to food logging, many modern nutrition apps offer meal-planning tools, recipe suggestions, and dietary adjustments based on users' nutritional preferences or restrictions. This functionality particularly benefits users looking to develop structured eating habits, increase meal variety, and meet specific health or fitness goals.



In future iterations, CIShrana plans to include advanced features, such as personalised meal-planning formulas tailored to individual health objectives, such as weight management, athletic performance, or chronic disease prevention.

The effectiveness of digital tools like nutrition apps in promoting healthier eating behaviors is well-documented. Studies have shown that consistent self-monitoring of food intake leads to improved weight management, better glycaemic control in individuals with diabetes, and greater adherence to dietary recommendations [5]. Nutrition apps provide immediate insight into caloric intake, macronutrient distribution, and micronutrient adequacy — data that would otherwise be difficult for individuals to estimate accurately without professional guidance. These apps empower individuals by offering real-time feedback and personalised insights, which is critical in improving long-term dietary adherence.

The adoption of nutrition apps offers significant benefits, both for individuals and healthcare professionals. Self-monitoring has been linked to improved diet quality, enhanced adherence to nutritional guidelines, and more successful management of chronic conditions. CIShrana is designed not only to appeal to individuals aiming to improve their lifestyle and nutrition but also to professionals such as dietitians and medical doctors. The app is already being embraced by a growing community, including students and professionals in nutrition and supplementation, particularly at the Faculty of Medical Sciences in Kragujevac. An active online forum further facilitates community engagement, allowing users to suggest new food items to be added to the app's database.

However, despite their benefits, nutrition apps are not without challenges. Some apps rely on incomplete or outdated nutritional information, leading to inaccuracies in dietary assessments. Additionally, continuously logging food intake may be perceived as tedious, potentially reducing long-term user adherence. To address these challenges, CIShrana has been designed to simplify the user experience while maintaining a high level of accuracy. Streamlined food tracking and automated, personalised dietary reports ensure that novice users and field experts can easily integrate the app into their routines. From a public health perspective, mobile nutrition apps represent a scalable solution for educating and empowering individuals to adopt healthier eating habits. Their integration into clinical workflows — particularly in preventive medicine, chronic disease management, and sports nutrition can reduce healthcare systems' burden

and improve long-term health outcomes. The future of mobile nutrition apps lies in their ability to integrate diverse data sources and offer real-time, personalised dietary guidance. As technology advances, these apps will likely become an essential component of a data-driven approach to public health.

In conclusion, while CIShrana was initially designed as a tool for synchronised swimmers, its broad potential as Serbia's first mobile nutrition diary underscores the power of sports science to inspire solutions that benefit the wider population. This application represents only the beginning of a much broader digital transformation in nutrition. As the field of nutrition continues to evolve, future versions of CIShrana will further personalise dietary guidance, integrate biometric data, and contribute to a more comprehensive, data-driven approach to nutrition management.

3. CONCLUSION

Nutrition applications promote healthier lifestyles by providing real-time dietary tracking, meal planning, and personalised nutritional guidance. While these apps have certain limitations, their ability to enhance dietary awareness and promote informed food choices is undeniable. The development of the first Serbian mobile nutrition diary marks a significant step toward localised, culturally sensitive digital health tools that bridge the gap between evidence-based nutrition and everyday practice. While global applications offer general frameworks for food logging, they often fail to reflect national dietary customs, leading to limited accuracy and poor user engagement in local populations. By incorporating national food items, meal structures, and portion sizes, our application addresses a long-standing need for a tailored approach in nutritional self-monitoring within the region.

CIShrana represents a significant milestone in Serbia's digital health landscape, offering the public localised, scientifically backed nutrition insights. As the application evolves, its impact on public health, athletic performance, and professional dietetics is expected to grow, reinforcing the importance of digital nutrition tools in contemporary wellness management.



4. ACKNOWLEDGEMENTS

We want to acknowledge the help of the Center for Sports Nutrition and Supplementation (CIS) and the Synchronized Swimming Federation of Serbia, who helped develop that idea into real everyday life and used by all people interested in their nutrition habits.

REFERENCES

- [1] W. H. O. (WHO), "Diet, nutrition and the prevention of chronic diseases. World Health Organ Tech Rep Ser. 2003;916," WHO Library Cataloguing-in-Publication Data, Geneva, 2002.
- [2] M. C. G. T. R. D. G. N. I. C. M. P. E. & G. R. Vitale, "Ultra-Processed Foods and Human Health: A Systematic Review and Meta-Analysis of Prospective Cohort Studies.," *Adv Nutr*, vol. 15, no. 1, p. <https://doi.org/10.1016/j.advnut.2023.09.009>, 2024.
- [3] W. M. S. J. M. J. H. D. S. S. Coughlin SS, "Smart-phone Applications for Promoting Healthy Diet and Nutrition: A Literature Review.," *Jacobs J Food Nutr.*, vol. 2 (3), no. 021, p. <https://pmc.ncbi.nlm.nih.gov>, 2015.
- [4] P. S. H. D. e. a. Paramastri R, "Comput Methods Programmes Biomed.," *Use of mobile applications to improve nutrition behaviour: A systematic review.*, vol. 192:105459, p. doi:10.1016/j.cmpb.2020.105459, 2020.
- [5] M. L. J. C. M. A. N. N. Ghelani DP, "Mobile Apps for Weight Management: A Review of the Latest Evidence to Inform Practice.," *Front Endocrinol (Lausanne)*. 2020;11:412. , vol. 11, no. 412, p. doi:10.3389/fendo.2020.00412. <https://doi.org/10.3389/fendo.2020.00412>, 2020.
- [6] N. K. K. J. K. H. Theodore Armand TP, "Applications of Artificial Intelligence, Machine Learning, and Deep Learning in Nutrition: A Systematic Review.," *Nutrients*, vol. 16(7), no. 1073, p. doi:10.3390/nu16071073, 2024.
- [7] H. W. A. J. C.-N. K. DiFilippo KN, "The use of mobile apps to improve nutrition outcomes: A systematic literature review.," *J Telemed Telecare*. 2015;21(5): , vol. 21, no. 5, pp. 243-253. doi:10.1177/1357633X15572203, 2015.
- [8] T. M. K. M. P. C. Payne JE, "Adherence to mobile-app-based dietary self-monitoring-Impact on weight loss in adults.," *Obes Sci Pract.*, vol. 28, no. 8(3), pp. 279-288. doi: 10.1002/osp4.566. , 2021.
- [9] H. L. N. M. M. e. a. Hauptmann, "Effects and challenges of using a nutrition assistance system: results of a long-term mixed-method study.," *User Model User-Adap Inter* , vol. 32, no. 923-975, pp. <https://doi.org/10.1007/s11257-021-09301-y>, 2022.
- [10] A. K. J. M. L. S. M. R. & S. T. Böhm, "Real-World Evidence of User Engagement With Mobile Health for Diabetes Management: Longitudinal Observational Study.," *JMIR Mhealth Uhealth.*, vol. 8, no. 11, PMID: 32975198; PMCID: PM, p. doi: 10.2196/22212., 2020.



PARENTS' ATTITUDES TOWARDS THE INTEGRATION OF INNOVATIVE TECHNOLOGIES IN PHYSICAL EDUCATION FOR PRESCHOOL CHILDREN

Petar Nikodijević¹,
[0009-0003-0373-4202]

Goran Vještica²,
[0009-0004-6414-1109]

Vladan Vođević³,
[0009-0006-9587-879X]

Aleksandar Gadžić^{4*}
[0000-0002-1611-6855]

¹Primary School Đura Jakšić,
Čuprija, Serbia

²Association for Preschool Sport,
Animation and Recreation,
Belgrade, Serbia

³Medical College of Applied Studies,
Belgrade, Serbia

⁴Singidunum University,
Belgrade, Serbia

Abstract:

In modern-era preschool settings, there is an emphasis on motor skills improvement because of their significance in the child's overall development. Basic motor skills such as walking, running, and jumping influence not only physical development but also cognitive, social, and emotional development. New technologies might increase both gross and fine motor skills by encouraging activity in children. Studies indicate that interactive tools, virtual environments, and educational apps improve both gross and fine motor skills. The primary objective of this study was to examine how well parents are familiar with and apply innovative technologies in the physical education of their children. As a secondary objective, we tried to identify any potential differences in the respondents' answers in relation to gender.

Study results show that most parents of both genders agree with the use of innovative technologies in working with their children, did not use applications or mobile platforms for physical education at home (without gender differences) and that they desire better collaboration with experts in this field.

Children today are encountering innovative technologies at an early age, and it is reasonable to incorporate these technologies into physical education programs as early as preschool. The time of pessimistic views on technology use in the context of physical education is behind us, and while we need to be aware of the risks, we also need to explore the opportunities for innovative technologies to change physical education in positive ways.

Keywords:

Children, Physical Activity, Mobile Apps, New Technologies.

INTRODUCTION

Motor development in preschool age is an important part of the overall development of a child. It enables the acquisition and refinement of basic physical abilities or motor skills such as walking, running, jumping, throwing, and catching. These skills form the foundation for later physical, cognitive, social, and emotional development. The development of fine motor skills, like precise movements of the hands and fingers, is also of fundamental importance, as it allows for performing everyday tasks such as drawing, writing, and dressing independently. Physical education in the preschool context stimulates children's physical development but also their emotional, social, and cognitive development. Well-designed physical education programs may enhance basic motor skills development

Correspondence:

Aleksandar Gadžić

e-mail:

agadzic@singidunum.ac.rs





and help children acquire values such as fair play and teamwork. Evidence strongly suggests that regular physical activity can prevent many health issues, including obesity and heart disease [1].

Motor skills have a great significance in the child's overall development. That is the reason why preschool settings of the modern era have an emphasis on their improvement. Basic motor skills such as walking, running, and jumping play a crucial role not only in physical development but also in cognitive, social, and emotional development. Inevitably, the emergence of new technologies caused adjustments in teaching methods. Today, we must admit that the use of electronic devices (smartphones, video games, tablets, etc.) has changed the way children interact and play. New methods increasingly incorporate interactive games and virtual realities to provide children with engaging and dynamic ways of learning. Studies indicate that interactive tools, virtual environments, and educational apps can improve gross and fine motor skills. These technologies might aid in the development of coordination, precision, balance, and overall physical fitness. At the same time, it can encourage additional engagement of children in physical activity, preventing obesity and other health issues [2]. Adequate use of these technologies can boost social skills such as collaboration, sharing, and conflict resolution. Furthermore, cognitive development can be stimulated through games that require problem-solving, logical thinking, and decision-making [3]. Nevertheless, it is important to maintain a balanced use of new technologies to avoid negative impacts on physical and social health [4]. Its effective use requires careful planning and integration with traditional educational practices.

The primary objective of this study was to examine how well parents are familiar with and apply innovative technologies in the physical education of their children. As a secondary objective, we tried to identify any potential differences in the respondents' answers in relation to gender.

2. MATERIAL AND METHOD

The sample in this study consisted of 30 parents of preschool-aged children. Of these, 16 parents were male, and 14 were female, mean age 43.5 years.

The survey conducted as part of this research was specifically designed for this purpose and was inspired by relevant scientific papers addressing the integration of technology into preschool education, particularly in the context of physical education. A significant influence

on the survey design includes works by several authors who explored how digital games, interactive technologies, and augmented reality can impact children's engagement and motor development [5] - [7]. The questionnaire consisted of 6 items that covered data on the use of applications and digital platforms in the physical education of a child, general learning or play, parents' collaboration with technology experts, and their belief in the positive impact of technology on motor development.

The above-stated items within the questionnaire were formulated as follows:

1. Have you ever used apps or online platforms for physical education at home with your child?
2. Have you noticed any changes in your child's motor skills that you can attribute to the use of technological tools for learning?
3. Would you like more collaboration with experts (educators, psychologists, IT specialists) when it comes to integrating technology into physical education?
4. Do you track and evaluate your child's progress in physical activities that involve technology?
5. Do you believe that technology can positively impact a child's motor development?
6. How do you assess the availability of information on how to use technology to support a child's motor development?

The data analysis was conducted in the form of the frequency distribution of responses (percentage distribution), and to compare positive and negative responses, as well as any potential differences between genders in the responses, the χ^2 test was applied to all variables.

3. RESULTS WITH DISCUSSION

As shown in Table 1, the distribution of responses to the posed questions is presented in the total number of positive and negative responses, and separately by the criterion of gender.

The results from Table 1 show that most parents, of both genders, agree with the use of innovative technologies in working with their children, that they monitor their children's progress in physical activity related to technology, and that they desire better collaboration with experts in this field (answers to questions 3, 4 and 5). On the other hand, the answer to question one had the least positive outcome (for both genders), and that is related to whether parents had ever used applications



or online platforms in the field of physical education at home. Answers to questions two and six had around one third of responses in an affirmative way indicating the somewhat lower effect of any changes in a child's motor skills that could be attributed to the use of technological tools and availability of information on how to use technology to support a child's motor development. The obtained results can also be interpreted in such a way that parents likely still lack sufficient technological literacy, as it is known that "digital natives" (younger generations) easily master certain digital devices for communication, while "digital immigrants" (the adult population) require specific communication and media education to successfully function in the modern world on a daily basis [8].

In a more detailed analysis of the data, we applied the χ^2 test to examine differences in responses overall by positive and negative answer choices, and then also differences between respondents by gender. The results of the χ^2 test showed that there is a statistically significant difference in the overall responses to question 1, which clearly indicates that most parents did not use applications or mobile platforms for physical education at home and there are no gender differences. These findings are in line with the previous studies that confirmed the absence of mobile applications or digital platforms

in parental involvement in children's physical activity [9], [10]. However, it is worth noting that there is also a significant number of studies in which it has been found that parents prefer not to use modern technology in the physical education of their children and instead opt for traditional approaches [11] - [13].

When considering the differences in responses based on the gender of the respondents, the results of the χ^2 test showed that there are no statistically significant differences in any of the responses. These results indicate that parents of both genders are equally aware of the advantages and disadvantages of using modern technologies in the physical education of their children. These results are also consistent with some earlier studies that have confirmed a consensus among parents regarding the advantages and disadvantages of using modern technologies in the physical education of their children [14], [15].

It has been confirmed that use of technology in the form of fitness apps, exergames, and wearable devices was found to be effective in motivating children to engage in physical activity. These technologies serve as support tools and as powerful motivators that offer tracking and feedback on their activity which can encourage children to engage in physical activity more actively [16].

Table 1. Distribution of answers to the questions

Question	Yes total	No total	Yes male	No male	Yes female	No female
1	13.3%	86.7%	6.7%	40.0%	6.7%	46.7%
2	30.0%	70.0%	15.3%	31.3%	16.7%	36.7%
3	70.0%	30.0%	32.3%	14.3%	35.7%	17.7%
4	70.0%	30.0%	30.3%	16.3%	36.7%	16.7%
5	70.0%	30.0%	29.3%	17.3%	32.7%	20.7%
6	33.3%	66.7%	16.7%	30.0%	16.7%	36.7%

Table 2. Results of χ^2 test

Question	χ^2 Value (Total)	df (Total)	p-value (Total)	χ^2 Value (Gender)	df (Gender)	p-value (Gender)
1	7.50	1	0.01	0.01	1	0.92
2	2.40	1	0.12	0.00	1	1.00
3	0.00	1	1.00	0.00	1	1.00
4	0.00	1	1.00	0.00	1	1.00
5	0.00	1	1.00	0.00	1	1.00
6	1.67	1	0.19	0.20	1	0.65



Figure 1. Children like to look up to their parents and use fitness trackers (AI generated image)

A review of the obtained results can be made from the perspective of the level of technological development and changes in the sphere of technologies that define both the relativity and variability of human technological education. In the context of human development as an individual, the variability of communication and technological education in adults reflects the changing educational needs in the domain of communications and media across different age groups. As people grow, their needs change, and this also applies to the needs in the realm of communication and technological literacy. In the earliest days, a child first communicates non-verbally, and then verbally, almost exclusively with their parents. Later, peers, friends, and colleagues enter their lives, and the ways and means of communicating with them are significantly different from those practiced within certain social roles, both in earlier and current stages. Furthermore, it becomes clear that from birth, as a person grows and matures, they encounter increasingly complex needs in the areas of communication and media interaction. Therefore, it is logical that communication and media-educational needs also change and multiply [8].

The primary school students daily use mobile technology, and that should be used in the activities in an advantageous way. Increasing children's awareness and capacity to engage in physical sports activities is one way to address this issue through a health promotion strategy. Previous studies have shown that elementary school students who utilize technology are better able to comprehend and implement basic physical exercises at school, on playgrounds, or at home, particularly when those activities are combined with engaging and varied elements [17]. As argued earlier, there is a need for new methods that would increasingly incorporate interactive

games and virtual realities to provide children with engaging and dynamic ways of learning. Some practical examples include numerous apps that combine games with physical activity, such as fitness challenges, location-based games, tracking fitness parameters, etc.

One of the best examples was the PLAY study that examined the feasibility and preliminary effectiveness of a mobile app, and parent-led curricula to promote fundamental motor skills proficiency for preschool children. The authors found it to be effective, with the potential for wide-scale dissemination to parents of preschoolers and to provide a model for the utilization of mobile apps to promote young children's motor skill development [18]. Another illustration is the smartphone app Jungle Gym, which was created by a team of writers to promote preschoolers' physical exercise. The software was designed to assist kids in practicing gross motor skills, improve vocabulary linked to movement, and provide parents and kids a chance to engage in movement-related conversation if they so desire. Parents who participated in the app prototype testing showed positive reactions to the app and thought it would be helpful in a range of scenarios. Preschoolers who participated in the test demonstrated a high level of engagement with the software; most of them stood and/or moved vigorously for the whole duration [19].

Furthermore, mobile apps found their use as a tool in fighting obesity in preschool children. Namely, the authors of MINISTOP 2.0 app claim that it has the potential to be implemented in primary child health care nationally (Sweden) and thereby reach many families who may benefit from it. They found it highly relevant considering that obesity is a major public health challenge globally [20].



4. CONCLUSION

As a result of the introduction of new technologies, teaching strategies ought to change. We must acknowledge that children's play and interaction these days have been altered by the usage of electronic gadgets (such as tablets, smartphones, and video games). Providing children with dynamic and interesting learning experiences, new approaches are increasingly using virtual reality and interactive games. By promoting activity in kids, new technology may improve their gross and fine motor skills.

This study's main goal was to find out how well parents understand and use cutting-edge technologies in their kids' physical education. Our secondary goal was to explore any possible gender-related discrepancies in the respondents' answers. Since children today are encountering innovative technologies at an increasingly early age it would be entirely reasonable to incorporate these technologies into physical education programs as early as preschool. It seems plausible to look at technology as a tool to support healthy lifestyles among preschool children. The time of pessimistic views on technology use in the context of physical education is behind us and while we need to be aware of the risks, we also need to explore the opportunities for innovative technologies to change physical education in positive ways. Additionally, it would be beneficial for educators or physical education teachers to make efforts and take steps toward educating parents about the positive aspects of such practices.

5. ACKNOWLEDGEMENTS

The authors would like to thank all parents who participated in this survey.

REFERENCES

- [1] B. Ziemowit, & S. Winiarski, "Correlations between angular velocities in selected joints and velocity of table tennis racket during topspin forehand and backhand," *J. Sports Sci. Med.*, vol. 17, no. 2, pp. 330-338, May. 2018.
- [2] S. Bozóki, L. Csató, & J. Temesi, "An application of incomplete pairwise comparison matrices for ranking top tennis players," *Eur. J. Oper. Res.*, vol. 248, no. 1, pp. 211-218, Feb. 2016, doi: 10.1016/j.ejor.2015.06.069
- [3] R. Vilhjalmsson, & G. Kristjansdottir, "Gender Differences in Physical Activity in Older Children and Adolescents: the central role of organized sport," *Soc. Sci. Med.*, vol. 56, no. 2, pp. 363-374, Jan. 2003, doi: 10.1016/s0277-9536(02)00042-4.
- [4] K. Ishii, A. Shibata, & M. Adachi, "Gender and grade differences in objectively measured physical activity and sedentary behavior patterns among Japanese children and adolescents: a cross-sectional study," *BMC Public Health*, 15, 1254, Dec. 2015, doi: 10.1186/s12889-015-2607-3
- [5] B. Bossavit, & I. Arnedillo-Sánchez, "Motion-based technology to support motor skills screening in developing children: A scoping review," *Comput. Methods Programs Biomed.*, vol. 240, no. 107715, Oct. 2023, doi: 10.1016/j.cmpb.2023.107715.
- [6] J. Zabatiero, L. Stone, D. McCormack, et al., "I Use Technologies Strategically with my Family Now: Practices That Parents Value to Promote Physical Activity in Young Children," *J. Phys. Act. Health*, vol. 21, no. 11, pp. 1121-1131, Sep. 2024, doi: 10.1123/jpah.2024-0317.
- [7] D. Arabiat, M.A. Jabery, S. Robinson, L. Whitehead, & E. Mörelius, "Interactive technology use and child development: A systematic review," *Child Care Health Dev.*, vol. 49, no. 4, pp. 679-715, Nov. 2022, doi: 10.1111/cch.13082.
- [8] N. Kačavenda-Radić, T. Nikolić-Maksić, & B. Ljujić, *Komponente kvaliteta komunikaciono-medijskog obrazovanja* Belgrade, Serbia: Filozofski fakultet Univerziteta u Beogradu, 2014.
- [9] C.L. Edwardson, & T. Gorely, "Parental influences on different types and intensities of physical activity in youth: A systematic review," *Pediatr. Exerc. Sci.*, vol. 22, no. 2, pp. 203-215, Nov. 2010, doi: 10.1016/j.psychsport.2010.05.001.
- [10] V.A. Goodyear, & A. Casey, "Innovation with change: Developing teachers' practices to support student-centered learning through physical education," *Sport Educ. Soc.*, vol. 20, no. 8, pp. 1075-1094, Sep. 2015, doi: 10.1080/17408989.2013.817012.
- [11] R.R. Pate, J.R. O'Neill, & J. Mitchell, "Measurement of Physical Activity in Preschool Children," *Med. Sci. Sport Exer.*, vol. 42, no. 3, pp. 508-512, Mar. 2010, doi: 10.1249/MSS.0b013e3181cea116.
- [12] T. Hinkley, J. Salmon, A.D. Okely, & S.G. Trost, "Correlates of Sedentary Behaviours in Preschool Children: A Review," *Int. J. Behav. Nutr. Phys. Act.*, vol. 7, no. 1, pp. 66, Sep. 2010, doi: 10.1186/1479-5868-7-66.
- [13] L.M. Vanderloo, "Screen-Viewing Among Preschoolers in Childcare: A Systematic Review," *BMC Public Health*, vol. 14, no. 1, pp. 205, Aug. 2014, doi: 10.1186/1471-2431-14-205.



- [14] A. Casey, V.A. Goodyear, & K.M. Armour, "Rethinking the relationship between pedagogy, technology, and learning in health and physical education," *Sport Educ. Soc.*, vol. 22, no. 2, pp. 288-304, Sep. 2017, doi: 10.1080/13573322.2016.1226792.
- [15] D. Lupton, & B. Williamson, "The datafied child: The dataveillance of children and the implications for their rights," *New Media Soc.*, vol. 19, no. 5, pp. 780-794, Jan. 2017, doi: 10.1177/1461444816686328
- [16] A. Rahayu, H. Setyawan, S. Sabariah, et al., "Strategies to increase physical activity in elementary school children in the digital age to support a healthy lifestyle," *Retos*, vol. 61, pp. 1410-1421, Dec. 2024.
- [17] T. Bopp, & M. Stollefson, "Practical and Ethical Considerations for Schools Using Social Media to Promote Physical Literacy in Youth," *Int. J. Environ. Res. Public Health*, vol. 17, no 4, pp. 1225, Feb. 2020, doi: 10.3390/ijerph17041225.
- [18] E.K. Webster, C.L. Kracht, Jr.R.L. Newton, R.A. Beyl, & A.E. Staiano, "Intervention to Improve Preschool Children's Fundamental Motor Skills: Protocol for a Parent-Focused, Mobile App-Based Comparative Effectiveness Trial," *JMIR Res. Protoc.*, vol. 9, no. 10, e19943, Oct. 2020, doi: 10.2196/19943.
- [19] M.L. McCloskey, D.A. Thompson, B. Chamberlin, L. Clark, S.L. Johnson, & L.L. Bellows, "Mobile Device Use Among Rural, Low-Income Families and the Feasibility of an App to Encourage Preschoolers Physical Activity: Qualitative Study," *JMIR Pediatr. Parent*, vol. 1, no. 2, e10858, Apr. 2018, doi: 10.2196/10858.
- [20] H. Henriksson, C. Alexandrou, P. Henriksson, et al., "MINISTOP 2.0: a smartphone app integrated in primary child health care to promote healthy diet and physical activity behaviours and prevent obesity in preschool-aged children: protocol for a hybrid design effectiveness-implementation study," *BMC Public Health*, vol. 20, 1756, Nov. 2020, doi: 10.1186/s12889-020-09808-w.



THE USAGE OF THE BLAZEPOD PLATFORM IN SPORTS

Vladislav Pavlović¹,
[0009-0007-4072-0407]

Vladan Marković^{2*},
[0009-0003-4708-0909]

Miloš Milošević²
[0000-0003-2592-451X]

¹Faculty of sport and psychology, TIMS,
Educons University,
Novi Sad, Serbia

²Singidunum University,
Belgrade, Serbia

Abstract:

The implementation of information technologies in sports has proven to be a fertile ground for innovation and improvement of sports performance. BlazePod platform is a tool for measuring and quantifying motor and cognitive parameters such as reaction time, agility, attention and concentration. In this research, different age groups, a total of 214 male and female trainees who practice real Aikido at the Martial Arts Center Gymnasium in Loznica, were tested. The trainees were tested using the BlazePod platform with three different reaction speed tests (simple, complex and distracting test). Statistically significant ($p < 0.05$) correlations of all variables were revealed, among which the negative relationship between age and impaired reaction time ($\rho = -0.88$) in women stands out. Numerous gender differences, as well as gender specificities of the researched relationships, were also discovered and described. The results confirm that the use of information technologies in sports, in this case, the BlazePod platform, gives the tested athletes valuable feedback in showing their motor and cognitive potential and helps in the creation of more effective training programs.

Keywords:

BlazePod Platform, Real Aikido, Reaction Speed, Cognitive Abilities, Test.

INTRODUCTION

In an era defined by rapid technological advancement, the implementation of information technology in sports has proven to be a fertile ground for innovation and performance improvement. The integration of digital tools and new methods into athlete training has changed the way athletes train, compete, and analyze their performance [1]. This synergy is particularly evident in the field of martial arts, where the fusion of traditional techniques with cutting-edge technology offers new avenues for skill development and cognitive enhancement [2]. Real Aikido, a dynamic and practical martial art, emphasizes adaptability, precision, and efficient use of the opponent's energy [3]. Real Aikido also emphasizes the principles of timing and focuses on redirecting the opponent's energy, which requires exceptional timing, coordination, and reaction time. Reaction time is the time between a stimulus and the first appropriate response [4].

Correspondence:

Vladan Marković

e-mail:

vladanmarkovic@singidunum.ac.rs





Real Aikido was founded by Ljubomir Vračarević, who wanted to remove religious elements and mystification from Aikido and emphasize its efficiency [5]. It is a Serbian martial art that emphasizes hitting the vital areas before applying leverage. It is used in the training of special forces such as the army, police and bodyguards [5].

The inclusion of technological means, such as the BlazePod platform, represents an opportunity to objectively measure and quantify key cognitive parameters such as reaction time, attention and concentration in different age groups practicing real aikido [6]. The use of information technologies in sports provides athletes with valuable feedback, thus shortening the time needed to acquire the necessary skills, which influences overall sports performance. The current training process is unthinkable without the use of multimedia sports support [7], while BlazePod, a reaction training system using light-based signals, has gained popularity in sports training for its ability to improve cognitive and physical performance [8].

Contemporary athlete training emphasizes a holistic approach, integrating physical fitness with cognitive training to optimize psycho-physical characteristics in various sports disciplines. Studying reaction time, attention and concentration in all age groups is crucial, as these mental abilities are fundamental to the successful execution of techniques in martial arts such as real Aikido. Martial arts training is associated with improvements in cognitive function in children and cognitive control in adults [9], and participation in programs in martial arts schools can lead to better efficiency, modulation of attention and excitability of the motor cortex, cognitive flexibility and the ability to anticipate changes in the environment [10].

Previous research has shown that sports interaction technology often targets different combinations of performance, engagement and learning [11]. Martial arts and sports, in particular, have the potential to positively impact these cognitive abilities, including reaction time, attention, and concentration.

This study aims to investigate the effectiveness of BlazePod technology through three tests (simple, complex, and choice reactions) to measure reaction time, attention, and concentration among real Aikido practitioners of different age groups. Simple reaction time requires only one response to one stimulus, while in reaction time, there is an alternative stimulus, and the subject makes a specific response to the presented stimulus [12]. The ultimate goal is to assess the capacity of these technologies to provide coaches and athletes

with valuable resources, improve athlete performance, and optimize training strategies, which will help them make optimal physical, technical, and tactical decisions [13]. By quantifying these cognitive parameters, we aim to provide insight into how real-world aikido training affects cognitive function across the lifespan and how technology can be used to optimize training strategies for practitioners of all ages. Although reaction time is the key component in martial arts, especially karate [14] because it requires a high level of execution of explosive techniques, research related to reaction time in this area is rare. Moreover, the existing literature regarding simple and choice reaction time in martial arts is contradictory [15]. Some studies suggest that martial arts practitioners possess superior sensory functions, while others indicate no significant differences. Studies suggest that karate training can improve sustained attention and reduce impulsivity, even in children with ADHD, and can also improve choice reaction time, which is important for explosive techniques. Athletes perform specific cognitive tasks faster and more accurately [16]. Increasing evidence suggests that physical activity can improve cognitive abilities in individuals of all ages.

This study is concerned with how to effectively measure simple reaction speed, choice reaction when there is more than one stimulus (for which the dominant factors are attention and concentration), and reaction time, i.e., concentration and attention when distracting factors are present.

2. METHOD

2.1. PARTICIPATION

A total of 214 participants took part in this research, all of whom are active members of the Real Aikido club "Gymnasium" in Loznica. The sample was stratified into five different age categories to account for developmental variations in cognitive and motor skills.

Preschool group (4-6 years old): 19 girls and 29 boys. The participants in this category held the following ranks: 10th kyu (30 participants), 9th kyu (6 participants), and 8th kyu (12 participants).

Younger pioneers (7-9 years old): 28 girls and 35 boys. Their ranks were distributed as follows: 10th kyu (12), 9th kyu (4), 8th kyu (15), 7th kyu (14), 6th kyu (11), 5th kyu (6), and 4th kyu (2).

Senior pioneers (10-12 years old): 30 girls and 17 boys. The distribution of ranks in this group was: 10th kyu (9), 8th kyu (6), 7th kyu (6), 6th kyu (6), 5th kyu (3), 4th kyu (3), 3rd kyu (1), and 2nd kyu (3).



Juniors (13-17 years old): 20 girls and 9 boys. Their ranks were: 10th kyu (10), 8th kyu (6), 7th kyu (4), 6th kyu (6), 5th kyu (5), 4th kyu (3), 3rd kyu (4), 2nd kyu (1), and 1st kyu (3).

Seniors and veterans (18+ years old): 9 females and 18 males. The participants in this group held the following ranks: 6th kyu (3), 5th kyu (1), 4th kyu (2), 3rd kyu (4), 1st kyu (1), 1st day (12), 2nd day (3), and 3rd day (2).

All participants had been training intensively for at least six months, with rankings ranging from children's yellow belt (10th kyu, beginner level) to advanced master title (black belt, 3rd day). The diverse experience levels among the participants allowed for a comprehensive analysis of the effects of BlazePod training across different skill levels.

2.2. EXPERIMENTAL PROCEDURES

This research seeks to comprehensively evaluate the reaction time of real Aikido trainees using the BlazePod platform, focusing in particular on simple reaction time, complex reaction time (a reflection of attention and concentration) and reaction time under the influence of distracting light and sound stimuli.

A single BlazePod device was used for simple reaction time assessment and data quantification. The device was programmed to emit blue light at unpredictable intervals, ranging from 0.5 to 4 seconds, thus preventing prediction and ensuring an honest response to the stimulus. Each test protocol lasted 30 seconds, providing enough data for a reliable analysis. The participants were tested individually, in the presence of an examiner, in a controlled environment to minimize any distractions during the test phase. The participants were instructed to stand at the table where the BlazePod was placed, maintaining a consistent hand position over the device. After the blue light was turned on, the participants were instructed to react as quickly as possible by touching the BlazePod to record their reaction time. Before the start of the actual test, each participant was allowed a practice trial in order to familiarize themselves with the test procedure and minimize any potential learning effects during the formal data collection phase.

For the complex assessment of reaction time, a task known as the "Clap Challenge", which is a protocol standardized by BlazePod for assessing attention and concentration, was implemented. In this protocol, four BlazePod devices are arranged linearly on a table, each separated by a distance of 20 centimeters. During the

30-second test, the device glowed in different colors, with three different lights flashing in an unpredictable order, which added cognitive load and required the participants' selective attention. The participants were instructed to respond to each color stimulus in a specific manner: the blue light was to be deactivated with the left hand, the red light with the right hand, and the green light required hand clapping before deactivation, introducing a motor coordination component to the task. Average reaction time, measured in milliseconds, was recorded upon completion of the test, which provided a quantitative measure of the participant's ability to process and respond to multiple stimuli under time constraints. The data was processed in BlazePod, which calculated the mean value in milliseconds after the completion of the protocol.

In order to simulate distraction during exercise in real-world situations, a third test was conducted that included distracting light and sound stimuli to mirror scenarios in which athletes must maintain focus amid ambient noise. In this test, the setup reflected a complex reaction time assessment, with four BlazePod devices arranged in a linear order on the table. During this test, the participants stood in front of a table with four BlazePods, which were arranged as in the previous test in order to reduce the effects of spatial familiarity. The participants were instructed to respond exclusively to a predefined color, while other colors served as distractors, which required increased attention control and cognitive inhibition. Additional complexity was provided by the audience present during this test, which introduced an element of social pressure and auditory distraction that could affect the participants' concentration and reaction time. This multifaceted approach enabled a comprehensive assessment of reaction time, attention and concentration at different levels of cognitive demands.

2.3. DATA ANALYSIS

Descriptive statistical analysis was conducted to describe and calculate the mean levels and dispersion of the examined variables in gender groups.

Instead of the variables that do not meet the criteria for the use of parametric statistics, this research employed testing differences between male and female groups using non-parametric the Mann-Whitney. To test the relationships between variables, we used Spearman's correlation. The level of statistical significance was set to $p < 0.05$ and $p < 0.01$.

Jamovi software (version 2.6.44) was used for statistical processing.



3. RESULTS

Table 1 presents the results of the descriptive analysis of the female subsample by age category.

Table 2 presents the results of the descriptive analysis of men in the subsample according to age category.

The significance of the difference in scores between men and women that can be seen in Tables 1 and 2 was

tested by analyzing variance using the non-parametric the Mann-Whitney, while differences between age groups were tested with non-parametric Krasakal-Wallis test, the results of which are shown in Table 3.

The results of the analysis of variance non-parametric the Mann-Whitney show no statistically significant ($p>0.05$) difference between women in any of the measured variables (Table 3). On the other hand, all differences between age groups are significant ($p<0.01$).

Table 1. Descriptive statistical analysis of female age categories. Values are presented as Mean $\pm$ Standard deviation

Cat	ps	yp	op	j	s
Age	5.3 $\pm$ 0.8	8.1 $\pm$ 0.8	10.7 $\pm$ 0.9	14.8 $\pm$ 1.1	29.8 $\pm$ 17.9
m lev	1.7 $\pm$ 0.9	3.8 $\pm$ 1.8	4.5 $\pm$ 2.8	7.8 $\pm$ 2	10.6 $\pm$ 1.9
tra ex	0.9 $\pm$ 0.5	2 $\pm$ 1	2.6 $\pm$ 2	4.8 $\pm$ 12.7	11.7 $\pm$ 7.2
com r	1213 $\pm$ 122	1036 $\pm$ 113	885 $\pm$ 157	734 $\pm$ 116	697 $\pm$ 235
dis r	943 $\pm$ 184	719 $\pm$ 184	561 $\pm$ 82.8	470 $\pm$ 146.2	503 $\pm$ 139
sim r	507 $\pm$ 94	416 $\pm$ 62	367 $\pm$ 119	351 $\pm$ 1134	312 $\pm$ 175

cat – category, m lev – mastery level, tra ex – training experience, com r – complex reaction [ms], dis r – disrupted reaction [ms], sim r – simple reaction [ms], ps - preschool, yp - young pioneers, op - older pioneers, j – juniors, s – seniors

Table 2. Descriptive statistical analysis of male age categories. Values are presented as Mean $\pm$ Standard deviation

Cat	ps	yp	op	j	s
Age	5.4 $\pm$ 0.8	8 $\pm$ 0.9	10.8 $\pm$ 0.9	14.8 $\pm$ 1.4	39.7 $\pm$ 19.2
m lev	1.6 $\pm$ 0.9	3.3 $\pm$ 1.6	5.1 $\pm$ 2.5	8.8 $\pm$ 1.3	10 $\pm$ 2.1
tra ex	0.8 $\pm$ 0.5	1.8 $\pm$ 1	2.7 $\pm$ 1.4	6.2 $\pm$ 3.2	9.5 $\pm$ 6.8
com r	1198 $\pm$ 131	1052 $\pm$ 146	863 $\pm$ 118	787 $\pm$ 166	780 $\pm$ 158
dis r	909 $\pm$ 188	718 $\pm$ 109	549 $\pm$ 55	485 $\pm$ 57	472 $\pm$ 88.6
sim r	482 $\pm$ 77	395 $\pm$ 89.3	334 $\pm$ 44.7	350 $\pm$ 87.7	269 $\pm$ 34.6

cat – category, m lev – mastery level, tra ex – training experience, com r – complex reaction [ms], dis r – disrupted reaction [ms], sim r – simple reaction [ms], ps - preschool, yp - young pioneers, op - older pioneers, j – juniors, s – seniors

Table 3. Non-parametric the Mann-Whitney test for the independent variable gender and non-parametric Krasakal-Wallis test for age group

	Mann-Whitney U	Wilcoxon W	Z	Significance		χ^2	Df	p	ϵ^2
Age	5147.5	11033.5	-1.28	0.2	Age	203	4	<0.01	0.955
Level	5312.5	11198.5	-0.92	0.36	Level	138	4	<0.01	0.648
training experience	5290	11176	-0.98	0.33	training experience	106	4	<0.01	0.498
complex reaction	4985.5	10656.5	-1.63	0.1	complex reaction	134	4	<0.01	0.628
disrupt reaction	5088.5	10759.5	-1.4	0.16	disrupt reaction	106	4	<0.01	0.496
simple reaction	5486.5	11372.5	-0.52	0.6	simple reaction	155	4	<0.01	0.729



Table 4. Spearman correlation for men and women

		age	mastery level	training experience
female	complex reaction	-0.779**	-0.697**	-0.626**
	disrupt reaction	-0.881**	-0.794**	-0.743**
	simple reaction	-0.782**	-0.663**	-0.584**
male	complex reaction	-0.367**	-0.709**	-0.548**
	disrupt reaction	-0.346**	-0.565**	-0.408**
	simple reaction	-0.347**	-0.337**	-0.293**

** $p < 0.01$

The correlation results for men and women are shown in Table 4.

The correlation results revealed the existence of a statistically significant negative correlation ($p < 0.01$) between all variables.

4. DISCUSSION

The results of this research show that BlazePod is an effective tool for measuring and improving reaction time, attention, and concentration in different age groups of real Aikido, and it is a highly motivating factor for practitioners. In addition to reaction speed, concentration and attention, with its help, it is possible to influence almost all motor and cognitive abilities. Attention and concentration are of utmost importance for sports that require quick reactions, and this fact should be taken into account when planning training. The inclusion of cognitive tasks in motor tasks can be more effective for improving cognitive abilities than separating mental and physical functions in training [16]. Although the statistical analysis in our research did not indicate significant differences between men and women in the average values of reaction time ($p > 0.05$), female athletes showed better results on average in tasks that require selective attention, focusing and mental control. In tests with more stimuli, female athletes showed better results than their male counterparts. However, they noted that this could be influenced by seniority in sports and the fact that women emphasize accuracy rather than speed [28]. Herein lies the advantage and the ability of the BlazePod system to use light stimuli in different colors for different exercises to introduce an element of cognitive engagement in physical training, which potentially leads to improvements in both cognitive and physical performance [17]. Both children and adolescents, through martial arts training, can improve perceptual and visual processes. The correlation analysis in our

work showed a significant correlation between age, skill level and age reaction time. Older and more experienced athletes displayed faster reaction times. Although there are studies which state that reaction speeds is relatively stable from 20-60, another study showed that reaction time in children is age-dependent. In contrast, the children's concentration was weaker than in adults [18]. In our research, younger respondents showed a slower reaction time as compared to older groups, which is particularly visible in complex reaction tests ($p < 0.01$). The most notable differences were observed between preschool children (4-6 years old) and senior pioneers (10-12 years old), as well as between younger pioneers (7-9 years old) and juniors (13-17 years old). These disparities can be attributed to neurophysiological maturation, cognitive development, and experience gained through prolonged sports engagement.

In the context of sports, reaction time is a critical component that determines the ability to perform maximally, especially in dynamic and fast-paced activities [19]. Athletes with faster reaction times are better able to predict and respond to rapidly changing stimuli, which gives them an advantage in competition [20]. This research showed that as the level of experience increased, the reaction time became shorter, which indicates the effect of training in improving cognitive-motor abilities. This fact also suggests that the continuous improvement of test results from the younger to the older ages is the product of a specific training procedure. Furthermore, taekwondo masters have confirmed that they have a better reaction speed to specific visual stimuli than the untrained [21], which suggests that long-term practice of martial arts can improve neuromotor excitability and reaction time of both large and small muscles. The tests of complex reactions and reactions to distracting stimuli showed that athletes with more experience better process and react to multiple stimuli at the same time but also show better resistance to distracting external



factors. Older age groups, as well as subjects with more extended training experience, had better results in tasks with increased cognitive complexity. Combat sports are precisely characterized by changing situations and the speed of decision-making.

5. CONCLUSION

The Conclusion section of the paper contains the closing words and compact statements that emanate as conclusions from the complete body of the text presented in the paper. It should be no longer than 300 words; this study confirms the value of the BlazePod platform as a tool for improving sports performance, especially in the domains of reaction time, attention and concentration. The research confirmed the importance of cognitive training in the context of sports and emphasized the importance of applying modern technologies in the training process. Although the results were not significantly affected by gender differences, age and experience variations, which are essential for creating effective training programs, were observed. The results show that this technology can be successfully applied in training programs to provide athletes with objective feedback on their cognitive abilities. Future studies could include a longer period of monitoring athletes, as well as the analysis of other cognitive and psychological parameters to understand better the more profound implications of using the BlazePod platform in sports. Analysis of psychological and cognitive factors in the training of athletes can provide deeper insights into the way specific motor skills develop during a sports career.

REFERENCES

- [1] A. Yongtawee i J. M. Woo, "The Influence of Gender, Sports Type and Training Experience on Cognitive Functions in Adolescent Athletes," *Exercise Science*, t. 26, br. 2, pp. 159-167, 2017.
- [2] A. Johnstone, Cognitive Changes Associated with Martial Arts Practice, United Kingdom: Bangor University, 2021.
- [3] M. Alesi, A. Bianco, J. Padulo, F. P. Vella, M. Petucci, A. Paoli, A. Palma i A. Pepi, "Motor and cognitive development: The role of karate," *Muscles, Ligaments and Tendons Journal*, t. 4, br. 2, pp. 114-120, 2014.
- [4] A. Johnstone i P. Mari-Beffa, "The effects of martial arts training on attentional networks in typical adults," *Frontiers in psychology*, t. 9, br. 80, 2018.
- [5] S. Milosavljevic, D. Matavulj i P. Lazarevic, "Producing acute pain by means of a strike as a relevant prerequisite for a successful realization of the lever technique in real aikido," *SPORT - Science & Practice*, t. 4, br. 1, pp. 5-13, 2014.
- [6] M. N. Anitha i V. R. Samuel, "Reaction time in sitting and standing postures among typical young adults," *Physiotherapy - The Journal of Indian Association of Physiotherapists*, t. 12, br. 2, pp. 58-62, 2018.
- [7] A. Ezhov, A. Zakharova i D. Kachalov, "Modern Light Sport Training Systems: Critical Analysis of Their Construction and Performance," in *icSPORTS*, t. Octobar, pp. 123-129, 2021.
- [8] P. Shi i X. Feng, "Motor skills and cognitive benefits in children and adolescents: Relationship, mechanism and perspectives," *Frontiers in Psychology*, t. 13, br. 1017825, 2022.
- [9] R. J. Kosinski, A literature review on reaction time., South Carolina: Clemson University, 2013.
- [10] R. Agrawal i P. Borkar, "Influence of martial art on self-efficacy and attention time span in adults: Systematic review.," *International Journal of Physical Education, Sport and Health*, t. 8, br. 3, pp. 151-157, 2021.
- [11] S. Mori, V. Ohtani i K. Imanaka, "Reaction times and anticipatory skills of karate athletes," *Human Movement Science*, t. 21, br. 2, pp. 213-230, 2002.
- [12] V. Pavlinovic, N. Foretic, N. Kovacevic, T. Galic, L. Lusic Kalcina, F. Mihanovic i T. Modric, "Cognitive and Motor Capacities Are Poorly Correlated with Agility in Early Pubertal Children: Gender-Stratified Analysis," *Applied Sciences*, t. 14, br. 8, p. 3148, 2024.
- [13] G. Giordano, M. Gomez-Lopez i M. Alesi, "Sports, executive functions and academic performance: A comparison between martial arts, team sports, and sedentary children.," *International Journal of Environmental Research and Public Health*, t. 18, br. 22, p. 11745, 2021.
- [14] B. Coskun, S. Kocak i N. Saritas, "The comparison of reaction times of karate athletes according to age, gender and status.," *Science, Movement and Health*, t. 14, br. 2, pp. 213-217, 2014.
- [15] H. Chaabene, Y. Hachana, E. Franchini, B. Mkapuer i K. Chamari, "Physical and physiological profile of elite karate athletes.," *Sports medicine*, t. 42, br. 10, pp. 829-843, 2012.
- [16] Y. Zekailiche, H. Bougrine, J. Khalfoun, N. Jebabli, R. Djerddi, F. S. Saoud i A. B. Abderrahman, "Impact of utilizing visual stimuli (BlazePod) on agility, vertical jump, and visual reaction time speed in under-19 Volleyball Players.," *Preprints.org*, t. 980, br. v1, 2024.



- [17] L. Gradinaru, S. Petracovshi, E. Bota, P. Merghes i M. Oravitan, "The effect of Blazepod Flash Reflex Training program on vertical jump in U15 female volleyball players," *Timisoara Physical Education and Rehabilitation Journal*, t. 16, br. 30, pp. 31-37, 2023.
- [18] K. Bucsuhaży i M. Semela, "Case study: Reaction time of children according to age," *Procedia Engineering*, t. 187, pp. 408-413., 2017.
- [19] V. Subramanyam, D. Jamuna i P. Kalavathi, "A study of reaction time and attention control among athletes of sports training centers of Sports Authority of India.," *International Journal of Sports Sciences and Fitness*, t. 6, br. 2, pp. 132-141, 2016.
- [20] V. E. Vasyuk, R. Yichen, V. V. Bakaev, D. A. Lukashovich i T. V. Karaseva, "Interactive sensor technologies in the training of athletes of the children's department of the student sports club," *Theory and Practice of Physical Culture*, t. 8, pp. 22-25, 2023.
- [21] P. Chung i G. Ng, "Taekwondo training improves the neuromotor excitability and reaction of large and small muscles.," *Physical therapy in sport*, t. 13, br. 3, pp. 163-169, 2012.



THE PROFOUND IMPACT OF THE MOBILE APPLICATION INDUSTRY ON FITNESS

Tamara Ratković*,
[0000-0001-9476-5423]

Uroš Jovanov
[0009-0000-6640-935X]

¹Singidunum University,
Belgrade, Serbia

Abstract:

The contemporary urban way of living is characterized by the diversification of lifestyles, basically relying on the influence of modern technologies, especially on the use of Mobile applications, which have become instrumental in supporting daily activities. The culture of healthy physical exercise, and consequently the development of the fitness industry, is one of the key denominators of the contemporary urban lifestyle. This phenomenon has transitioned from the realm of health to that of a balanced lifestyle, which has been commercialized and, as a result, integrated into the broader culture of consumerism. The paper aims to present a program framework for the multifaceted utilization of digital mobile applications in the fitness industry about the diverse requirements of individuals in contemporary urban daily life. The paper establishes a cause-and-effect relationship between contemporary urban life and the fitness industry, highlighting four primary spheres of utilization for mobile applications. These include facilitating faster and more comprehensive communication between fitness service providers and users, supporting flexible lifestyle timetables, adapting to lifestyle changes related to age, gender, specific needs, and aspirations, and promoting quality control and progress monitoring of individually set goals.

Keywords:

Contemporary Lifestyle, Urban Everyday Life, Digital Technology, Fitness Industry.

INTRODUCTION

In relating the overall digital context of the contemporary urban way of living as a founding ground and the backbone for the utilization of Mobile applications in all spheres of life, in which the culture of healthy physical exercise is one of the key aspects, this paper presents a programmatic framework for the multifaceted introduction of digital mobile applications in the fitness industry.

Regarding the omnipresent individualization of urban society, it is widely accepted that Mobile applications make everyday life easier and less stressful. On the other hand, as products of high consumerism, they represent the appearance of belonging to the ideals of a cosmopolitan community. In balancing one's needs and aspirations, a wide network of mobile platforms has developed to support a diversity of urban ways of life, e.g., the fitness industry.

Correspondence:

Tamara Ratković

e-mail:

tratkovic@singidunum.ac.rs





The aim of this paper is to establish a cause-and-effect relationship between contemporary urban life and the fitness industry, highlighting key spheres of utilization for mobile applications.

The paper is structured in three parts, the first being the overall outline of contemporary urban lifestyle and the culture of healthy physical exercise, the second is the interlacing and establishing a cause-and-effect relationship between urban life and the fitness industry, and the last one is the introduction of Mobile applications with key spheres being presented through worldwide best practices in the field of the fitness industry.

Results show that even though the possibilities of utilizing mobile applications are without borders or limits, it is necessary to provide awareness of the knowledge not perceivable through artificial intelligence, securing a priority for individual encounters over virtual ones for the benefit of overall well-being.

2. CONTEMPORARY URBAN LIFESTYLE

Urban everyday life can be defined as a distinct way of life that differs significantly from suburban and rural lifestyles. The "cosmopolitan lifestyle" or "New York lifestyle" is a distinctive way of life that is intrinsically linked to the concept of the city as the primary facilitator of communication, integrating the diverse activities and contents of its citizens. The hallmark of the urban lifestyle is its heterogeneity, characterized by a vast array of contents and activities that cater to a diverse range of inclinations and routines among its inhabitants. This diversity is attributable to the substantial population of the city, which fosters a milieu conducive to the fulfillment of a vast array of lifestyles. [1] [2]. As [3] asserts, lifestyles represent the unique competencies of cities. Lifestyle can be understood as an expression of diverse individual tastes and preferences, and it is related to the free choice that the urban way of life implies. However, the urban way of life, characterized by its diversity of styles, is deeply entrenched in the quotidian routine of life, encompassing both private and business routines, as well as activities that must be accommodated within a 24-hour timeframe. [4].

Wirth's seminal work on urbanism delineates four fundamental characteristics of the modern urban environment that persist in their relevance today [5]. These are: a) a high degree of mobility that leads to the depersonalization of human relationships and their formalization, b) being burdened by a large number and diverse nature of activities, both professional and in the

sphere of leisure time, striving for the rationalization of space and time frame, c) the permeation of activities of a private and professional life, which is in cooperation with demographic changes, structure of the urban population and their needs and habits, and as a result of all previously mentioned, today in the 21st- century d) individualization of all spheres of everydayness. In the context of contemporary urban everyday life in the first half of the 21st century, the previous characteristics can be interpreted as follows:

- The integration of remote work and in-person professional activities within the same time frame;
- A flexible allocation of time among work, leisure, personal, and family domains;
- The evolution of personal interests and activities outside the domestic environment into a component of everyday social engagement;
- The acceleration and unpredictability of daily activities, coupled with high levels of mobility on both a daily and weekly basis;
- The integration of diverse domains such as leisure, education, work, and entertainment into a cohesive, fluid entity devoid of a precisely delineated schedule or durations;

In the context of quotidian life, individuals adhere to stylistic determinants that reflect their individual preferences and habits. This adherence complicates the determination of temporal, durational, and spatial parameters that define quotidian activities. Consequently, individuals perpetually adjust to these parameters. Their development schedule is frequently hard to establish in advance due to the constant interaction with the city and its inhabitants [6]. Consequently, activities that facilitate shared experiences between individuals and the city allocate minimal time for personal and mental well-being, which must often adapt to the prevailing rhythm of preceding activities. The aforementioned way of life necessitates a state of preparedness and stability, both physical and mental, a considerable degree of energy, and adherence to the standards of a good and quality life. This, in essence, entails a healthy way of life. In response to the challenges posed by urban living, there has been a notable development of a culture of healthy exercise. This cultural phenomenon has emerged as a critical dimension of urban everyday life, wielding a growing influence over the past two decades. The complexity of this cultural shift has transcended the domain of health, extending into the realm of lifestyle. It has evolved into a social norm, a value, and a behavioural paradigm.



Incorporated into routine activities, it prioritizes health and longevity, thereby gaining significant traction over the past decade, particularly due to technological advancements that facilitate real-time monitoring of its benefits [7]. The concept of the "Quantified Self", a contemporary phenomenon characterized by personalized and quantitative monitoring of achievements in the domains of health and active living through the utilization of smartwatch technology, has gained prominence. This technological tool quantifies and facilitates the realization of healthy habits promoted by the wellness culture [8], ranging from exercise (cloud health status) to quality sleep, nutrition, water consumption, and weight and body composition management, among others (human-centric platform). These recommendations are customized to align with an individual's age, specific needs, and lifestyle habits, ensuring a personalized approach to health enhancement [9].

In such circumstances, the culture of healthy exercise becomes one of the fundamental prerequisites of urban life and a mandatory component of daily activities. The prevailing 20th-century paradigm regarding exercise culture, which advocates for incorporating regular exercise at least two to three times per week in clearly delineated sessions in a designated location for 45 to 90 minutes, becomes impractical in the context of unpredictable daily life. In recent years, the culture of healthy exercise has evolved, shifting from a focus on structured exercise regimens to a more holistic approach that incorporates balanced nutrition and a well-managed daily routine. This shift has led to the emergence of diverse exercise modalities, such as Yoga, Zumba, Aqua-cycling, the Keto diet, and the Paleo diet, among others, all of which are tailored to individual lifestyles. Exercise has become commercialized and institutionalized, catering to individual needs. In the context of the urban lifestyle, the fitness industry, shaped in this manner, must respond to the diverse demands of daily life. In the context of the previous changes imposed by daily life in the 21st century, the following questions emerge.

- First, how should individuals adapt their exercise routines to the unpredictable daily rhythm of private and professional activities and their mobility?
- Secondly, how to integrate healthy exercise activities into an already complex mix of private, professional, educational, and leisure activities is of particular interest.
- Furthermore, how can diverse lifestyles be put into a cohesive set of routines that align with the principles of healthy exercise?

- Furthermore, in which way to develop metrics to assess and monitor the progress of individuals utilizing fitness services?
- Finally, how to enhance the flexibility of communication between users and fitness service providers to ensure the effective delivery of services and the satisfaction of all parties involved?

The aforementioned inquiries constitute specific requirements, primarily with regard to the flexibility and adaptability of the program, its implementation modalities, scheduling, the time and duration of health-promoting exercise activities, and the adaptation to a high degree of user mobility. Additionally, there is a need for the networking of several different activities and content that supports a holistic approach to wellness culture. This culture, in addition to an active culture of healthy exercise, encompasses several other activities and contents that support the previous ones.

3. DIGITAL TECHNOLOGIES AND THE CULTURE OF MOBILE APPLICATIONS

Digital technologies have reconfigured the value of authentic communication among individuals, as well as between individuals and municipal functions and content. Consequently, the paradigm of communication shifted from the physical city to the virtual realm [10]; [11]. The advent of the pandemic has further exacerbated this trend, leading to a pronounced shift in human interaction patterns, particularly in terms of physical interaction and temporal dynamics. The advent of virtual cyber communications and virtual spaces and activities has reconceptualised the relationship between humans and space [12], and consequently, the relationship to everyday routines, and also to the culture of healthy exercise and consequently to the fitness industry, building a new framework in the literature known as mFitt, i.e. Fitness Apps [13]. Visual experience has emerged as the predominant quality of spatial experience, while the perpetual possibility of communication has become the quintessential quality of temporal experience. Digital technologies, facilitated by mobile applications, have emerged as instrumental in providing these experiences within this domain [14]; [15]. The integration of technology into domestic life has become a prevalent aspect of contemporary existence, accompanied by the utilization of applications for a wide range of activities. The integration of Culture m APPS (Culture of Mobile Applications) into our lives has become pervasive. These applications have become so pervasive that their absence



would significantly hinder our daily lives. These applications have been designed with the intention for use across all spheres of life, especially for daily routines.

mApp (Mobile application) is rapidly entering the Fitness industry and through mFitt (Mobile Fitness) establishes a standard of response to the aforementioned needs of the culture of healthy exercise in relation to the modern urban lifestyle. It has been incorporated in several satisfactory ways:

1. The possibility of visiting a fitness club in several locations and even in several cities, depending on where the user is located;
2. Flexibility of the visit time and adaptability of the provided services in relation to the current needs and wishes of the client, as well as his health condition (stress level, exhaustion, way of spending time during work, available time, exercise intensity, etc.);
3. Adapting exercises to lifestyle (remote access, several times a week shorter, fewer times a week longer, variety of activities, and the like.
4. Networking of several different contents that follow a healthy way of exercising and living as well as their permanent monitoring.

3.1. EXAMPLES OF GOOD PRACTICE IN THE APPLICATION OF MFITT

In metropolitan regions, digital platforms and smart devices have emerged as pivotal intermediaries between service recipients and providers, thus fostering the proliferation of mFitt applications. In alignment with the aforementioned requisites, this segment of our study will henceforth present a compendium of exemplary practices presenting the extensive accessibility of mobile applications within the domain of fitness.

3.2. SUPPORTING FLEXIBILITY BETWEEN FITNESS USERS AND PROVIDERS

In support of the possibility of visiting a fitness club in multiple locations through applications that provide support in communication between users and service providers, a large number of platforms have been developed in the last decade that allow visiting a larger number of different fitness clubs through a unique system. This development has facilitated the provision of fitness services, allowing users to access a diverse array of activities and content across multiple locations, times, and fitness brands. Additionally, it has enabled fitness

clubs to offer their programs and services in a more flexible, efficient, and competitive manner. Prominent platforms developed in this domain include Fitpass, ClassPass, Gympass, Urban Sports Club, WellnessLiving, Mindbody, and MoveGB, among others.

FitPass is perhaps the most popular in our region, available in all major cities of Serbia. This subscription provides users with access to a large number of gyms, sports facilities, and swimming pools. To gain entry, users simply need to scan the application or QR code, which grants them access to the facility and its amenities (<https://fitpass.rs>). Comparable platforms include the Urban Sports Club platform, which is present in the European Union (<https://urbansportsclub.com/en>), the WellnessLiving platform, and the MoveGB platform, which is recognized in Great Britain (<https://get.movegb.com/>). The ClassPass platform employs a more comprehensive approach, offering access to fitness and wellness centers through a credit system that facilitates the booking of desired sessions and treatments. ClassPass has gained notable popularity in the United States, Europe, and Asia (<https://classpass.com/>). Gympass, for instance, is designed for corporate fitness programs, offering corporate employees access to fitness centers and online training through a single subscription (<https://wellhub.com/en-us/employees/>).

3.3. ENHANCING COMMUNICATION BETWEEN SPECIFIC FITNESS BRAND AND THEIR USERS

To ensure the flexibility and adaptability of providing multiple services within a single brand, digital platforms and mobile applications are being developed. These platforms facilitate faster, more accessible, more open, and simpler communication between the client and the chosen brand. This communication occurs through several different programs within one package of services, as well as through the flexibility of their mutual combinability. This mode of connection finds frequent application in modern fitness services, leveraging digital tools, social networks, and interactive platforms.

A notable example is AFC Studio Star, which utilizes the GoFitness application to enhance communication with clients and streamline appointment scheduling. This application empowers clients to create personalized schedules that align with their commitments and adjust them dynamically according to their individual rhythms. The client is empowered to select the preferred time, day, and program of interest (Pilates, corrective exercises, individual class, or BODYART). A notable



aspect of the platform is its flexibility, allowing clients to exercise autonomy in scheduling, location, and program selection. This adaptability is further exemplified by the comprehensive range of options, including Pilates, corrective exercises, personalized classes, and BODYART, as detailed on the AFC Studio Star website.

4. AI PERSONAL SUPPORT

In order to facilitate adaptation to rapid changes in needs and methods of exercise, as well as the implementation of exercise programs with assessment and monitoring of individualized services according to clients and their desired goals and health conditions, starting from sensor data logging, activity recognition, to on-top skill assessment, using the phone's built-in sensors, is essential. The contemporary generation of people utilizes smartphones, which are equipped with a plethora of sensing, communication, and computing capabilities. These devices offer an ideal platform for replicating some of the repetitive services of personal trainers in a way that is accessible to and affordable for virtually everyone [16]. This integration of mobile applications into everyday life is a significant development, fostering a closer and more intimate relationship between technology and society. This integration into daily life is accompanied by the potential for numerous benefits.

A notable holistic example is the UP150 concept [17], which focuses on corporations and provides employees with a permanent program of all-day shorter health exercises that encourage a healthy lifestyle. This program, which is installed in the office space, uses a

wide variety of tools and devices to remind employees to regularly hydrate, take short physical activity breaks, and play relaxing anti-stress music. The comprehensive program, which is part of work duties, also provides the satisfaction of personal needs, both health-related, emotional, and motivational.

Numerous applications are available to assist in the proper implementation of exercises. For instance, WellToDo (<https://www.welltodoglobal.com/post/virtuagym-launches-ai-powered-fitness-coach/>) and similar programs employ AI to suggest a set of exercises and automatically adjust the program through progress monitoring, ensuring the correct execution of exercises through a video connection. Other applications, such as Alfa-ai and StayFitCentral (<https://www.alfa-ai.com>; <https://stayfitcentral.com/endura-the-only-ai-workout-coach-youll-ever-buy/>), employ similar AI-powered methods to suggest exercises and adjust programs based on a user's health status and several quantifiable health measurements. The primary function of this mFitt application is to provide assistance in the execution of exercises, motivation during and prior to exercise, and support in daily routines that promote physical activity. Additionally, it is noteworthy that the application can adapt exercises to suit the individual abilities and preferences of clients.

In the context of Serbia, a notable illustration is provided by the Dr. FeelGood initiative, which, through the Zona Zdravlja application, offers a program that enables the personalization of the speed, time, and location of implementation, along with accompanying services that focus on the promotion of healthy eating habits and behaviors.

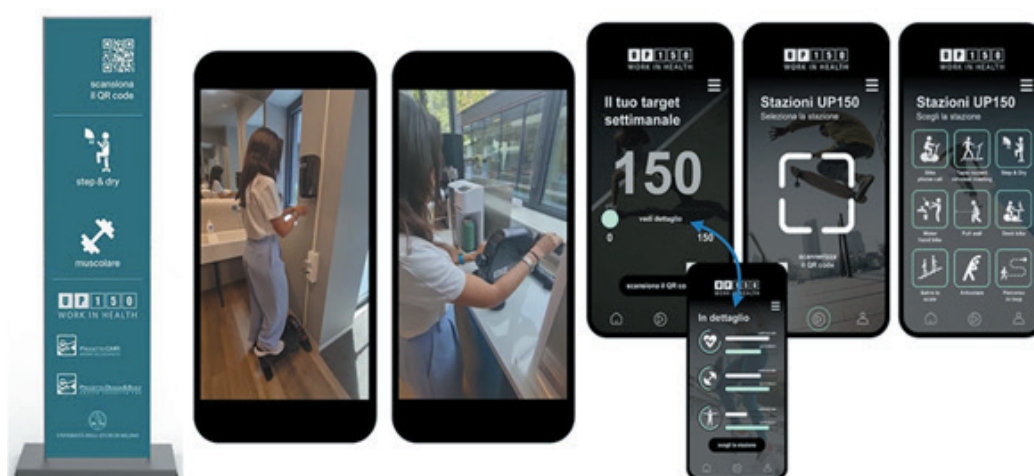


Figure 1. Example of architectural changes and setting for the UP150 office concept, and interaction with the UP150 App [18]



4.1. SMART MOBILE FITNESS AND WELLNESS APP

These platforms empower users by allowing them to customize their program and exercise system, dietary habits, and overall health. These applications facilitate the tracking of progress, the provision of real-time recommendations, and the integration of various components of a healthy lifestyle, sleep, stress level, and other factors into the daily routine. Daily activation control utilizes mobile phones and smart watches to monitor activities. CloudFit is a prominent example of a smart mobile platform for fitness and wellness, offering holistic health programs with personalized exercise, nutrition, and sleep plans. This application was developed by former athletes and includes a comprehensive wellness program that addresses the physical, psychological, emotional, and health aspects of well-being.

Additionally, there are devices designed for data collection, such as the GOQi wearable device, which gathers information without undergoing further analysis or providing activities and content. Other examples include Samsung Health and Apple Health, which facilitate the recording and management of fitness activities such as running, cycling, and swimming. The integration of the Galaxy Watch with the aforementioned application enables users to exercise more efficiently.

These platforms constitute an integral component of a progressively sophisticated ecosystem of smart fitness and wellness applications that leverage advanced technologies to enhance users' health and overall well-being. These applications' capacity to customize exercise regimens according to individual requirements and to deliver real-time assistance renders them a highly beneficial resource for individuals seeking to adopt a healthier and more active lifestyle.

5. DISCUSSION AND CONCLUSION

The present day is characterized by a profound interconnection between artificial intelligence (AI) technologies and urban everyday life. The fitness industry, undergoing rapid development, is susceptible to adjustments and transformations in its internal and external structure and content through the implementation of mFitt applications. The mFitt has the capacity to address and fulfill the demands of contemporary life, encompassing flexible scheduling and location, customized and adaptable programs and content, and continuous expansion of services. Drawing from illustrative cases of best practices, it can be concluded that the utilization

of mobile applications is pervasive. These applications, which are branded and part of the fitness industry, are primarily aimed at the urban population and are driven by commercial interests. The primary advantage of these applications lies in their ability to promote the integration of fitness into all aspects of life, thereby enhancing overall well-being and facilitating the attainment of a healthy and prosperous lifestyle. However, it is evident that certain aspects of a healthy physical exercise routine, particularly those that necessitate a sophisticated body of knowledge that trainers can impart to users of services, cannot be fully addressed by technological services or digital applications. Consequently, it is imperative to employ these technologies judiciously and to prioritize educational programs that serve not only as lifelong learning opportunities for educators and service providers but also for service users who, in the context of contemporary life, may not be able to promptly seek in-person professional assistance and guidance. The heterogeneity of the product range can generate an effect that is antithetical to the prevailing trend of fitness consumerism. This phenomenon does not stem from a contemporary and authentic demand for health benefits that emerge from an urban, predominantly sedentary lifestyle. Rather, it manifests as a form of leisure activity, a fleeting trend in its own right. Consequently, there is a pressing need to educate users in all domains of the fitness industry to mitigate the misconceptions surrounding the substitution of personal services for technologically advanced and artificial solutions to the greatest extent possible.

REFERENCES

- [1] D. B. Clarke, *The consumer society and postmodern city*, London: Routledge, 2003.
- [2] R. Fishman, "Urbanity and suburbanity: Rethinking the Burbs," in *American Quarterly*, 1994, pp. 35-49.
- [3] S. Zukin, "Diversity and standardization in space of consumption," *Urban lifestyles*, vol. Urban studies 35, pp. 825-839, 1998.
- [4] T. Hagerstrand, "Diorama, path and project," *Tijdschrift voor Economische en Sociale Geografie*, vol. 73, pp. 323-339, 1982.
- [5] L. Wirth, "Urbanism as a way of life," *American Journal of Sociology*, vol. vol.44 no.1, pp. 1-24, 1938.
- [6] J. Malpass, "The place of landscape concepts, contexts, studies.," *Mass. MIT press*, no. Cambridge, 2011.



- [7] L. Robertson, *The rise of Fitness Culture*, The Seraphim, 2021.
- [8] E. M. B. Marschke, "Creating a wellness culture through human resources," *Journal of Physical Education and Sports Management.*, vol. Vol.1, no. 1, pp. 61-80, 2014 . .
- [9] K. Dyer, "Quantified Body," *Digital Culture & Society* 2, vol. 2, p. 161, 2016.
- [10] M. Poster, "Everyday (Virtual) Life," *New Literary History*, vol. 33(4), pp. 743-760, 2002.
- [11] J. & E. O. Eniola, "Navigating New Realities:The Impact of Virtual Communities on Modern Lifestyles," 2025.
- [12] L. Rajendran, "Everyday landscape and meaning in urban living," *An International Review of the Humanities and Social Sciences.*, no. AGATHOS, pp. 120-131, 2013.
- [13] W. H. J. Z. J. S. Jiang Hu, "Examining the impacts of fitness app features on user well-being," *Information & Management*, vol. 60, no. 5, 2023.
- [14] S. G.-F. J.-P. M. Angosto, "A systematic review of intention to use fitness apps," *Humanit Soc Sci Commun*, vol. 10, 2023.
- [15] M. R. A. H. H. A. J. M. K. A. & A. R. Soliman, "The impact of mobile phone fitness applications on the level of physical fitness and psychological well-being during covid-19," *Journal of Education and Health Promotion*, vol. 11, p. 299, 2022.
- [16] A. S. T. P. P. R. M. Kranz, "The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices.," *Pervasive and Mobile Computing.*, vol. Volume 9, no. Issue 2, 2013.
- [17] P. L. I. R. S. M. M. a. Pietro Luigi Invernizzi, "The UP150: A Multifactorial Environmental Intervention to Promote Employee Physical and Mental Well-Bein," *Environmental Research and Public Health*, vol. 19 (3), 2021.
- [18] M. R. A. H. H. A. J. M. K. A. & A. R. Soliman, "The impact of mobile phone fitness applications on the level of physical fitness and psychological well-being during covid-19: The case of university students.," *Journal of Education and Health Promotion*, vol. 11, p. 299, 2022.



ISOMETRIC STRENGTH DIFFERENCES IN THE ATHLETIC SHOULDER TEST IN FEMALE VOLLEYBALL PLAYERS

Damjana V. Čabarkapa^{1*},
[0000-0003-4680-0313]

Vladimir Banković¹,
[0009-0000-5133-4383]

Nenad Trunić¹,
[0000-0001-5339-7695]

Dimitrije Čabarkapa²
[0000-0001-9912-3251]

¹Singidunum University,
Belgrade, Serbia

²University of Kansas,
Lawrence, USA

Abstract:

Volleyball is a sport that requires both lower- and upper-body power, particularly during explosive actions such as attacking and serving. Interestingly, while assessments like the countermovement jump are widely used to evaluate athletes' neuromuscular performance, upper-body testing protocols remain less commonly implemented in this sport. In response to this gap, a fairly new upper-body test known as the Athletic Shoulder (ASH) test has begun to be incorporated into evaluations within this cohort, aiming to measure isometric shoulder strength across three positions: I (180°), Y (135°), and T (90°). The purpose of this study was to examine positional differences in isometric strength during the ASH test among elite female volleyball players. Thirteen professional athletes competing at the SuperLeague level performed three maximal isometric contractions in each position using a force plate system. Significant differences were observed across positions ($p < 0.001$). Force output in the I position was significantly higher than both Y ($p < 0.001$, $g = 1.793$) and T ($p < 0.001$, $g = 2.880$), with Y also exceeding the T ($p = 0.015$, $g = 1.439$). These findings emphasize the importance of incorporating the ASH test into regular athlete monitoring protocols in volleyball, as they help coaches identify potential weaknesses, develop individualized training programs, and better understand athletes' overall functional capacity.

Keywords:

Overhead Sports, Performance Monitoring, Force Plate, Upper-Body Strength.

INTRODUCTION

Volleyball is a team sport that involves high-intensity actions (jumping, changes of direction) interspersed with brief periods of rest [1, 2, 3, 4, 5]. Over recent years, with exponential technological growth, various assessments have been included in the training and monitoring process to evaluate athletes' progress, optimize their performance, as well as reduce the risk of injury [2, 6, 7]. These assessments often include technologies such as force plate systems, motion capture systems, and wearable inertial measurement units that allow practitioners to obtain objective data on neuromuscular performance, movement quality, and levels of fatigue [5, 7, 8, 9].

Correspondence:

Damjana V. Čabarkapa

e-mail:

dcabarkapa@sindgidunum.ac.rs





Due to the nature of the sport, which involves frequent explosive lower-limb actions such as jumping, the countermovement jump (CMJ) is one of the most commonly utilized assessments in volleyball [2, 10, 11]. Data obtained from this assessment is then used by sports scientists and strength and conditioning coaches to analyze neuromuscular performance characteristics, track training adaptations, as well as guide further training decisions. However, relying solely on a single performance test may provide an incomplete or potentially misleading picture of an athlete's overall physical capabilities. Therefore, practitioners started incorporating additional tests such as isometric mid-thigh pull (IMTP), drop jump, handgrip strength test, etc., in order to obtain a broader profile of an athlete's strength capabilities and overall performance [12, 13, 14]. However, despite the growing range of performance tests that are currently being implemented within the volleyball population, relatively few are specifically designed to assess upper-body strength and power characteristics. This can be considered an important limitation, considering that volleyball is an overhead sport where upper-body function plays a critical role in performance, as well as injury prevention [15].

The handgrip strength test, assessed via the dynamometer, has been widely used as a simple yet reliable indicator of athletes' upper body strength [14, 16]. In addition, previous research has indicated that this test has a significant correlation with performance outcomes in various sports, including serving velocity in volleyball [17] and throwing velocity in handball [14, 18], highlighting its importance in overhead sports. However, in 2018, a new upper-body assessment was introduced in the rugby setting [19]. Ashworth and colleagues developed an isometric upper-body strength test called the Athletic Shoulder (ASH) test, designed to evaluate shoulder strength in overhead positions that closely mimic sport-specific demands and that are most vulnerable to injury [19]. However, despite the growing popularity and use of the ASH test in overhead sports, there is still a significant lack of research exploring the use of this specific assessment in elite female athletes, particularly in volleyball. Additionally, limited data exist on how isometric strength differs between the I, Y, and T testing positions, each position targeting different components of the shoulder musculature. Therefore, the aim of the present study was to evaluate positional differences in force output during the ASH test among elite female volleyball players.

2. METHODS

2.1. PARTICIPANTS

Thirteen professional female volleyball players ($\bar{X} \pm SD$; age=22.1 $\pm$ 3.5 years; height=184.5 $\pm$ 6.7 cm; body mass=76.3 $\pm$ 10.5 kg) competing at the SuperLeague level of competition participated in the present investigation. All procedures were approved by the Institutional Review Board (IRB) prior to data collection, and all participants were informed about the study's purpose and procedures, and written informed consent was obtained before participation.

2.2. PROCEDURES

Each athlete performed three maximal isometric contractions lasting 3 to 5 seconds with their dominant arm, with a 20-second rest between each attempt. The ASH test was conducted in a prone position, with the shoulder positioned at three consecutive angles of abduction (180°, 135°, and 90°). A one-minute rest was provided between each angle to minimize fatigue. The angles were pre-determined using a goniometer and properly marked to ensure consistency across trials (Figure 1).

To minimize extraneous movements, participants were instructed to maintain a neutral spine and avoid excessive trunk rotation or compensation. The dominant arm was identified based on self-reported handedness. A force plate system (Kinvent, Montpellier, France) (Figure 2) was used to assess maximal isometric force production, with data collected at 1000 Hz. The highest recorded maximal force value during the contraction was obtained for each position and used for the analysis.

Before testing, participants underwent a standardized warm-up, including submaximal contractions at each testing position, to ensure familiarity with the protocol. A practice trial was performed at each angle before data collection to reduce the learning effects.

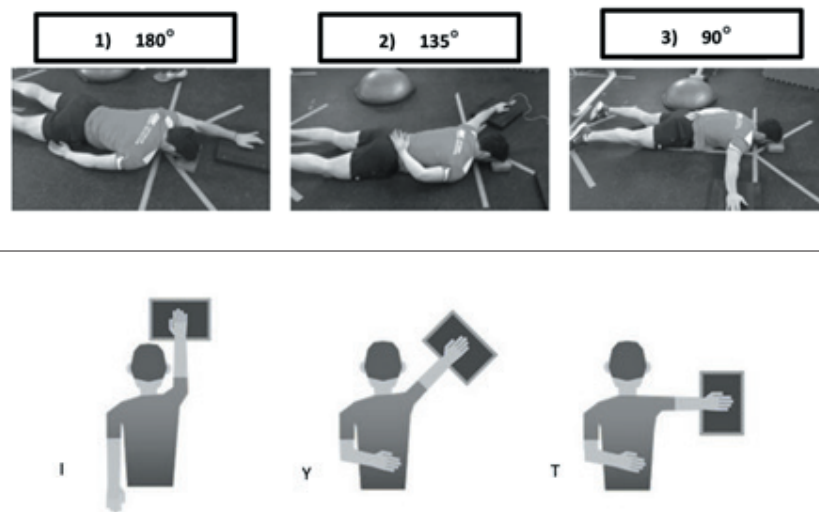


Figure 1. Graphical representation of each position during the Athletic Shoulder Test. [19]



Figure 2. Representation of the force platforms used in the present investigation

2.3. STATISTICAL ANALYSES

Descriptive statistics, means and standard deviations ($\bar{x} \pm SD$) were calculated for maximal force in each isometric position. Shapiro-Wilk's test corroborated that the assumption of normality was not violated for any of the dependent variables examined in the present study. A one-way analysis of variance (ANOVA) with Tukey post-hoc was used to determine statistically significant differences in maximal isometric force production between the *I*, *Y*, and *T* positions. Given the relatively small sample size ($n=13$), Hedge's g was used to calculate effect sizes (i.e., $g=0.2$ is a small effect, $g=0.5$ is a moderate effect, and $g>0.8$ is a large effect [20]). Statistical significance was set *a priori* to $p<0.05$. All statistical analyses were completed with SPSS (Version 26.0; IBM Corp., Armonk, NY, USA).

3. RESULTS

Significant differences in maximal force production during the ASH test were observed across the examined isometric positions ($F_{[2,23]}=27, p<0.001$) (Figure 3). Specifically, athletes generated the greatest force in the *I* position, which was significantly higher than both the *Y* ($p<0.001, g=1.793$) and *T* ($p<0.001, g=2.880$) positions. Force production in the *Y* position was also significantly greater than in the *T* position ($p=0.015, g=1.439$). On average, force production in the *I* position was approximately 24.81% greater than in the *Y* position and 46.90% greater than in the *T* position, while the *Y* position produced 17.70% more force than the *T* position.

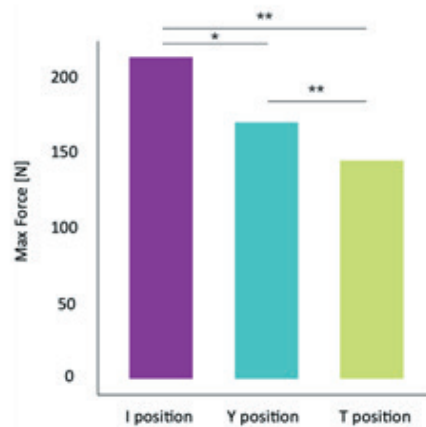


Figure 3. Graphical representation of maximal force production during the Athletic Shoulder Test in each isometric position. (*) – significantly different when compared to the Y position; (**) – significantly different when compared to the T position

4. DISCUSSION

The primary aim of this study was to examine positional differences in maximal isometric force production during the ASH test in elite female volleyball players. The results revealed significantly greater force output in the I position compared to both the Y and T positions, with the lowest force production observed in the T position. The aforementioned differences may be attributed to the muscle recruitment differences at different shoulder angles. For example, Joseph et al. [21] reported that the Y and T positions activate key periscapular and rotator cuff muscles, such as the lower trapezius and serratus anterior, to a higher degree, with lower trapezius activation exceeding 80% MVIC in both positions. While the Y position also demonstrated significantly greater serratus anterior activity, the T position showed reduced activation in the middle infraspinatus [21]. These findings, combined with the observed lower force outputs in the Y and T positions in the present study, suggest that these angles rely more heavily on smaller stabilizing muscles, in contrast to the I position, which likely allows greater contribution from larger prime movers and offers a mechanical advantage due to the arm being aligned with the line of force application.

Furthermore, similar observations to the ones in the present investigation have been noted by Ashworth and colleagues [19]. Specifically, during the ASH test, authors reported higher isometric force values in the I position among male rugby players (151.5 ± 51.7 – 155.5 ± 49.2 dominant arm; 144.9 ± 48.9 – 143.4 ± 47.3 non-dominant arm) when compared to the Y position (132.8 ± 41.1 – 133.3 ± 39.8 dominant arm; 125.2 ± 37.5

– 132.7 ± 38.7 non-dominant arm) and T position (123.8 ± 32.8 – 125.4 ± 35.6 dominant arm; 118.6 ± 29.9 – 121.8 ± 33.1 non-dominant arm) [19]. This is of critical importance for sports such as volleyball and rugby, as a great number of injuries in these sports occur during overhead actions such as attacking or throwing [19]. For example, in volleyball, the I position closely replicates the shoulder angle observed during the terminal phase of an attack or serve, where maximal shoulder flexion, external rotation, and force transmission are required. The ability to produce greater forces during this position may not only reflect an athlete's performance potential but also serve as a protective factor against common overhead-related injuries such as rotator cuff strains or labral tears [22, 23]. In addition, it is interesting to note that volleyball athletes seem to produce greater force during the ASH test, regardless of shoulder positioning, when compared to rugby athletes [19]. This may be attributed to the sport-specific demands since volleyball places greater emphasis on repeated, high-velocity overhead movements, and rugby involves more horizontal or forward-pressing movements (e.g., carrying the ball, tackling). Also, differences in training regimens that athletes went through, as well as the time point of the data collection during the season, can further explain these discrepancies. In addition, volleyball athletes were undergoing the ASH test frequently during their regular testing battery with their strength and conditioning coach thus, their familiarization with the assessment may have allowed them to obtain better results [24].



Lastly, while the findings of the present investigation provide valuable information regarding the isometric upper-body strength of elite female volleyball players, this study is not without limitations. The sample of athletes participating in the study was relatively limited in size and contained only female subjects. Hence, future research should examine if these findings are sex-specific as well as if they remain applicable to different levels of volleyball competition. In addition, it may be beneficial for practitioners to examine how the isometric upper-body strength differs across different positions in volleyball (e.g., outside hitter, middle blocker, libero).

5. CONCLUSION

In conclusion, the results of the present investigation highlight significant differences in isometric force production across the I, Y, and T positions of the ASH test in elite female volleyball players, with the I position eliciting the greatest force output. These findings emphasize the importance of incorporating the ASH test into regular athlete monitoring protocols in sports such as volleyball, as they help coaches to better identify potential weaknesses, develop individualized training programs, support injury prevention strategies, and obtain a better understanding of athletes' overall functional capacity. Also, this investigation is of critical importance as it contributes to the lack of elite female-athlete research and offers normative values for this specific group, which may help coaches and athletic trainers develop better training and recovery strategies.

REFERENCES

- [1] D. V. Cabarkapa, D. Cabarkapa, A. C. Fry, W. M. Shay and G. G. Downey, "Kinetic and kinematic characteristics of setting motions in female volleyball players," *Biomechanics*, vol. 2, no. 4, pp. 538-546, 2022. doi: 10.3390/biomechanics2040042
- [2] D. V. Cabarkapa, D. Cabarkapa, S. M. Whiting and A. C. Fry, "Fatigue-induced neuromuscular performance changes in professional male volleyball players," *Sports*, vol. 11, no. 6, p. 2023, 2023. doi: 10.3390/sports11060120
- [3] T. Gabbett, B. Georgieff, S. Anderson, B. Cotton, D. Savovic and L. Nicholson, "Changes in skill and physical fitness following training in talent-identified volleyball players," *Journal of Strength and Conditioning Research*, vol. 20, no. 1, pp. 29-35, 2006. doi: 10.1519/R-16814.1
- [4] R. Lidor and G. Ziv, "Physical characteristics and physiological attributes of adolescent volleyball players-A review," *Pediatric Exercise Science*, vol. 22, no. 1, pp. 114-134, 2010. doi: 10.1123/pes.22.1.114
- [5] T. G. Vlandes and T. Readdy, "Using microsensor technology to quantify match demands in collegiate women's volleyball," *Journal of Strength and Conditioning Research*, vol. 31, no. 12, pp. 3266-3278, 2017. doi: 10.1519/JSC.0000000000002208
- [6] A. J. Coutts, S. Crowcroft and T. Kempton, "Developing athlete monitoring systems - Theoretical basis and practical applications," in *Recovery and well-being in sport and exercise*, Routledge, 2021, p. 15. doi: 10.4324/9781003258117-3
- [7] G. Beckham, T. Suchomel and S. Mitzuguchi, "Force plate use in performance monitoring and sport science testing," *New Studies in Athletics*, vol. 29, no. 3, pp. 25-37, 2014.
- [8] F. Damji, K. MacDonald, M. A. Hunt, J. Taunton and A. Scott, "Using the VERT wearable device to monitor jumping loads in elite volleyball athletes," *PLoS One*, vol. 16, no. 1, p. e0245299, 2021. doi:10.1371/journal.pone.0245299
- [9] D. Cabarkapa, D. V. Cabarkapa, J. D. Miller, T. T. Templin, L. L. Frazer, D. P. Nicoletta and A. C. Fry, "Biomechanical characteristics of proficient free-throw shooters-markerless motion capture analysis," *Frontiers in sports and active living*, vol. 5, p. 1208915, 2023. doi: 10.3389/fspor.2023.1208915
- [10] T. Sattler, D. Sekulic, V. Hadzic, O. Uljevic and E. Dervisevic, "Vertical jumping tests in volleyball: reliability, validity, and playing-position specifics," *Journal of Strength and Conditioning Research*, vol. 26, no. 6, pp. 1532-1538, 2012. doi: 10.1519/JSC.0b013e318234e838
- [11] P. Schons, R. G. Da Rosa, G. Fischer, G. P. Berriel, C. G. Fritsch, F. Y. Nakamura, B. M. Baroni and L. A. Peyere-Tartaruga, "The relationship between strength asymmetries and jumping performance in professional volleyball players," *Sports Biomechanics*, vol. 18, no. 5, pp. 515-526, 2019. doi: 10.5114/jhk/193480
- [12] D. B. Martinez, "Consideration for power and capacity in volleyball vertical jump performance," *Strength and Conditioning Journal*, vol. 39, no. 4, pp. 36-48, 2017. doi: 10.1519/SSC.0000000000000297
- [13] A. Khanna and S. Koley, "Comparison of anthropometric profile and handgrip strength between inter-university volleyball players and a reference group," *Biomedical Human Kinetics*, vol. 12, pp. 82-90, 2020. doi: 10.2478/bhk-2020-0011
- [14] J. Cronin, T. Lawton, N. Harris, A. Kilding and D. T. McMaster, "A brief review of handgrip strength and sport performance," *Journal of Strength and Conditioning Research*, vol. 31, no. 11, pp. 3187-3217, 2017. doi: 10.1519/JSC.0000000000002149



- [15] A. Tsoukos, S. Drikos, L. E. Brown, K. Sotiropoulos, P. Veligekas and G. C. Bogdanis, "Upper and lower body power are strong predictors for selection of male junior national volleyball team players," *Journal of Strength and Conditioning Research*, vol. 33, no. 10, pp. 2760-2767, 2019. doi: 10.1519/JSC.0000000000002472
- [16] D. V. Cabarkapa, D. Cabarkapa, V. Bankovic and A. C. Fry, "Relationship between handgrip strength and vertical jump performance in Olympic female volleyball athletes," *Journal of Strength and Conditioning Research*, p. in press, 2025.
- [17] D. R. Melrose, F. J. Spaniol, M. E. Bohling and R. A. Bonnette, "Physiological and performance characteristics of adolescent club volleyball players," *Journal of Strength and Conditioning Research*, vol. 21, no. 2, pp. 481-486, 2007. doi: 10.1519/R-19185.1
- [18] L. Tsakalou, Z. Kotsampouikidou, M. Papa and I. Zapartidis, "Handgrip strength and ball velocity of young male and female handball players," *Journal of Physical Education and Sport*, vol. 15, no. 4, pp. 800-804, 2015. doi: 10.7752/jpes.2015.04122
- [19] B. Ashworth, P. Hogben, N. Singh, L. Tulloch and D. D. Cohen, "The Athletic Shoulder (ASH) test: reliability of a novel upper body isometric strength test in elite rugby players," *BMJ Open Sport & Exercise Medicine*, vol. 4, no. 1, p. e000365, 2018. doi: :10.1136/bmjsem-2018-000365
- [20] L. V. Hedges, "Estimation of effect size from a series of independent experiments," in *Psychological Bulletin*, Washington, D.C., *American Psychological Association*, 1982, pp. 490-499.
- [21] R. Joseph, T. Alenabi, T. Lulic and C. R. Dickerson, "Activation of supraspinatus and infraspinatus partitions and periscapular musculature during rehabilitative elastic resistance exercise," *American Journal of Physical Medicine and Rehabilitation*, vol. 98, no. 5, pp. 407-415, 2019. doi: 10.1097/PHM.0000000000001116
- [22] A. Aliprandi, S. Sdao, P. M. Cannao, Y. J. Khattak, S. Longo, L. M. Sconifienza and F. Sardanelli, "Imaging of shoulder pain in overhead throwing athletes," *Sport Sciences for Health*, vol. 9, pp. 81-88, 2013. doi: 10.1007/s11332-013-0151-z
- [23] E. Seminati and A. E. Minetti, "Overuse in volleyball training/practice: A review on shoulder and spine-related injuries," *Sports and Exercise Medicine and Health*, vol. 13, no. 6, pp. 732-743, 2013. doi:10.1080/17461391.2013.773090
- [24] D. Drake, R. Kennedy and E. Wallace, "Familiarization, validity and smallest detectable difference of the isometric squat test in evaluating maximal strength," *Journal of Sports Sciences*, vol. 36, no. 18, pp. 2087-2095, 2018. doi:10.1080/02640414.2018.1436857



EVALUATION OF TIME MEASUREMENT METHODS – A REVIEW

Srđan Marković^{1*},
[0000-0001-6009-3817]

Sladana Rakić^{1,2}
[0000-0002-2222-2507]

¹Singidunum University,
Belgrade, Serbia

²Fiziogroup
Belgrade, Serbia

Abstract:

Accurate sprint time measurement is essential for assessing athletes' speed capabilities and informing training decisions. This review paper examines the characteristics of various timing methods, including electronic timing systems (such as photocells and digital setups) and manual methods (like stopwatches), with a focus on their validity, reliability, accuracy, and practicality across different sports contexts. The analysis includes ten scientific papers published between 2020 and 2025 that compared measurement methods in both laboratory and field settings. The findings indicate that electronic systems significantly reduce human error, providing greater accuracy and consistency in measurements compared to stopwatches. Nevertheless, manual methods remain valuable in resource-limited environments, particularly when standardised protocols are followed. Additionally, mobile applications emerge as a promising alternative, especially when complemented by software analysis. In conclusion, the choice of measurement method should be guided by the test objectives, the level of expertise, and the available resources. For situations requiring high precision, the use of electronic systems is recommended.

Keywords:

Sprint, Photocells, Stopwatch, Reliability, Validity.

INTRODUCTION

Accurate and reliable measurement of running time is crucial in sports, particularly when assessing, selecting, and enhancing athletes' performance. Time measurement in various disciplines, such as sprints, agility tests, and specific sports evaluations, serves as the foundation for making important decisions for coaches, researchers, and athletes alike. Traditionally, stopwatches have been the primary tool for timing because of their practicality, affordability, and ease of use. However, with advancements in technology, photocells are becoming more common and are favoured for their higher accuracy and reliability [1]. Despite this trend, stopwatches are still frequently used, especially in circumstances where cost, availability, or practicality are significant concerns.

To better understand the importance of timing in sports, it is essential to define several key terms: validity, accuracy, and precision. Validity refers to the degree to which a particular measurement method accurately measures what it claims to measure.

Correspondence:

Srđan Marković

e-mail:

smarkovic@singidunum.ac.rs





In the context of sports performance, validity is critical because it directly impacts the interpretation of results and decision-making, which can have significant consequences for athlete training, selection, or readiness assessment [2]. Accuracy indicates how close a measured result is to the actual value of a performance, such as running time. High accuracy is crucial, as even minor deviations can affect the evaluation of athletic abilities and the effectiveness of training programs. Reliability refers to the consistency of repeated measurements taken under identical conditions and indicates the dependability of the measurement instrument or method [3].

Understanding these terms is essential for effectively evaluating sports performance. In practice, stopwatches that require manual activation and stopping often have limitations regarding accuracy and reliability, as their results are significantly influenced by the reaction time of the person operating them [4]. As a result, stopwatches are susceptible to both systematic and random errors, which can lead to significant variations in the measurement of the same event by different individuals or even by the same person at different times. Photocells work by cutting infrared light beams, providing automated start and stop times without requiring human intervention. This dramatically reduces the impact of human error, enhancing the accuracy and precision of measurements [1].

In recent years, numerous studies have compared these two methods to identify apparent differences in their metric characteristics. Understanding these differences is crucial because selecting the appropriate measurement method can significantly impact the quality of the information obtained, which in turn affects the decisions made based on that information. Aside from scientific research, the choice of method also depends on practical factors, such as the cost of the device, its market availability, the technical complexity of operating the device, and the specific conditions under which the measurements are taken [5].

New technologies and systems, including mobile applications, video analysis, and GPS-based systems, have been developed to enhance timing capabilities. [6, 7, 8]. These innovations aim to combine the convenience of stopwatches with the precision and accuracy of photocells. Despite these advancements, photocells are still considered the gold standard in most research studies where high precision and reliable results are essential [9].

This paper offers a comprehensive overview of research findings from the past five years, providing valuable insights into the advantages and disadvantages, as well as the practical implications, of using photocells and stopwatches. By analysing previous studies, we can offer clear recommendations for the application of these tools in various sports contexts. Additionally, this review aims to benefit the research community by consolidating existing knowledge and identifying areas where further research is needed, particularly in the introduction of new technologies.

Given the complexity and significance of this topic, this review aims to provide a thorough, systematic, and detailed overview of the current state of time measurement techniques using photocells and stopwatches. This paper is intended for coaches and researchers who need guidance in selecting the most effective and reliable methods for measuring sports performance. A clear understanding of these methods and their characteristics will enhance the training process, improve athlete selection, and increase the validity of conclusions drawn from test results.

2. METHODS

This review utilised a systematic analysis of relevant scientific literature published in the last five years (2020-2025) to provide a comprehensive and reliable overview of current knowledge. This analysis aimed to identify and compare the metric characteristics of running time measurement using photocells and stopwatches, with a specific focus on the validity, accuracy, and precision of the instruments employed.

The literature search was conducted using the following Google Scholar electronic database. The keywords employed in the search included *stopwatch reliability*, *photocell accuracy* and *sprint time measurement*.

Only peer-reviewed scientific papers that contained empirical research on sprint timing or similar movements in athletes were considered. Papers focused solely on laboratory testing without applications in a sports context were excluded, as were review papers that did not provide specific data on the metric characteristics of the instruments.

Several papers were selected that met all the inclusion criteria. Each paper was analysed in terms of:

- types of measuring instrument (stopwatch, photocell, application),
- sizes and characteristics of the respondents (gender, age, sport),



measurement method and start position (reactive or automated),

displayed values for reliability (e.g. ICC, CV%), statistical validation and level of precision.

Special attention is paid to the differences in validity between methods involving the human factor (stopwatches) and automated systems (photocells, applications), as well as the effects of variability between measurement repetitions. Potential measurement biases, including the meter's *learning effect* and the influence of ambient conditions on the measurement, were also analysed.

This methodology enabled an objective and structured analysis of contemporary literature, laying the foundations for the following sections of the paper, where the results and implications of the findings will be presented and compared in detail.

3. RESULTS AND DISCUSSION

An analysis of contemporary research published between 2020 and 2025 enabled the identification of key differences in metric characteristics between running time measurements using stopwatches, apps and photocells. The papers focused on parameters such as accuracy, precision, and validity of measurements in different sports contexts (Table 1).

3.1. VALIDITY AND RELIABILITY OF MOBILE APPS

Chen et al. [1] investigated the validity and reliability of the mobile app *COD Timer* in comparison to hand-held stopwatches for measuring direction change performance (Figure 1). Their findings indicated that the *COD Timer* application demonstrated high validity and reliability, whereas the manual stopwatches exhibited greater measurement errors, highlighting their limited effectiveness for precise measurements.

Table 1. Comparison of the metrics of stopwatches and photocells

Characteristic	Stopwatch	Photocells
Accuracy	Low	High
Precision	Variable	High
Reliability - ICC	0.70-0.85	>0.95
Human error influence	High	Low
Practicality	High	Medium
Equipment price	Low	High

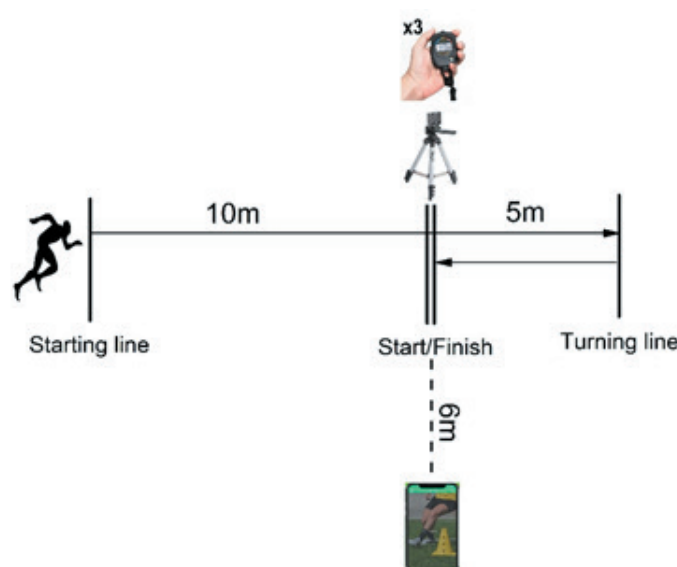


Figure 1. Different ways of measuring time [1]



Uysal et al. [5] assessed the validity, reliability, and sensitivity of mobile applications for estimating turn speed. Thirty students participated in the tests using six different devices, which included apps and handheld stopwatches. The study found that some apps delivered results comparable to professional systems, while the handheld stopwatches were less reliable.

It can be concluded that while mobile applications and stopwatch-based time measurements are feasible, using mobile applications is preferable for ensuring the validity and reliability of results.

3.2. COMPARISON OF DIFFERENT MEASURING DEVICES

Altmann et al. [9] studied the criterion validity of various devices used to assess maximum sprint speed in soccer players. Thirty elite young soccer players participated in 30-meter sprint tests utilising radar, photocells, a magnetic timing system, and GNSS devices. The results indicated that photocells and magnetic systems delivered the most accurate measurements, whereas GNSS devices showed greater variability.

Zajac et al. [10] assessed the validity and inter-examiner reliability of manual measurements for maximum sprint speed. The study involved 18 amateur runners, and the results revealed significant discrepancies between manual measurements and electronic systems. This highlights the need for caution when relying on manual stopwatches.

Based on previous statements, we recommend using electronic systems, specifically photocells, for timing running tests, especially when measuring athletes' maximum speed. While all the other measuring devices mentioned are effective, electronic systems offer the highest level of accuracy.

3.3. APPLICATION OF NEW TECHNOLOGIES IN SPRINT MEASUREMENT

Skujytė et al. [11] investigated the reliability and validity of the *Alex7* device for measuring sprint performance under conditions with added load and assistance. The results showed that *Alex7* provides reliable measurements but consistently overestimates run times compared to standard devices, indicating the need for calibration when using this device.

Dewanti and Hermawan [2] compared the measurement of running speed using photocells and manual stopwatches to reduce human error. The study showed that photocells provided more accurate results compared to manual stopwatches, which were more prone to measurement errors.

Integrating stopwatches into the testing process is the most feasible and straightforward approach; however, it is essential to consider that measurement error, resulting from the human factor, is present. To what extent it will depend on various conditions. Indeed, once again, a commitment to electronic devices, but with the caution of monitoring the calibration of the given instruments.

3.4. COMBINING DIFFERENT MEASUREMENT METHODS

Van den Tillaar et al. [12] compared the mechanical parameters of sprinting measured by photocells and a laser device. Thirty-four female handball players performed 30-meter sprints, with the results showing that both devices provide similar measures, but with certain differences that can be significant depending on the context of the application.

Mamo [3] highlighted the superiority of photocells in measuring sprint performance compared to manual stopwatches, emphasising that even minimal measurement errors can significantly affect the evaluation of athletes.

Marco-Contreras et al. [13] analysed the validity and reliability of the *Photo Finish*® mobile app for measuring sprint times. The study showed that the application provides accurate and reliable results, with an error of 0.09 seconds compared to photocells, which makes it suitable for use in various sports contexts.

Moreno-Azze et al. [4] evaluated the validity and reliability of the iOS mobile application *Fitnessmeter* for measuring change of direction performance. Forty-six adolescent soccer players participated in the study, and the results showed a strong correlation between the app and standard photocells, with no significant differences in times, suggesting that the app is a valid performance measurement tool.

These findings indicate that, although photocells are superior in terms of accuracy and reliability, manual stopwatches may still have their applications in situations where resources are limited or when a quick assessment of performance is required. However, it is necessary to be aware of their limitations and interpret the results with caution.



4. CONCLUSION

Analysis of these studies reveals that photocells and certain mobile applications are more reliable for sprint timing than manual stopwatches. Manual stopwatches are prone to human error and are less accurate, particularly in situations that require high precision. Mobile applications, such as *COD Timer* and *Photo Finish*®, have proven to be effective alternatives, offering accurate results with greater convenience and accessibility.

However, using mobile applications necessitates clearly defined measurement protocols and a certain level of technical training for users to ensure accuracy and reliability. Additionally, variations in smartphone models and the recording speed of cameras can affect the quality of the results, indicating that these applications cannot yet fully replace standardised systems, such as photocells or laser timing devices.

It's important to recognise that the effectiveness of these technologies depends on the specific context and available resources. While photocells are commonly used in professional sports environments, mobile applications can be beneficial in school or recreational settings where resources may be limited. Furthermore, factors such as user experience, test conditions, and the technical specifications of the equipment should be considered when selecting a measurement method.

In addition, several studies suggest that combining multiple methods, such as using a mobile application with parallel photocell measurement, can contribute to validation and facilitate a more accurate interpretation of the results. The introduction of automated data analysis and processing through software platforms also opens up space for more precise monitoring of athletes' progress over time.

The choice of sprint timing method should be guided by a balance between required precision and available resources, taking into account the specific demands of the sporting context and performance evaluation objectives. The potential for further development lies in standardising the use of applications and integrating digital systems that offer both measurement and analysis within the same ecosystem. However, it is important to note that the application of these technologies depends on the specific context and resources. While photocells are standard in professional sports environments, mobile applications can be useful in school or recreational settings where resources are limited. Additionally, factors such as user experience, test conditions, and technical specifications of the equipment should be considered when selecting a measurement method.

REFERENCES

- [1] Z. Chen, C. Bian, K. Liao, C. Bishop and Y. Li, "Validity and Reliability of a Phone App and Stopwatch for the Measurement of 505 Change Direction Performance: A Test-Retest Study Design," *Frontiers in Physiology*, vol. 12, pp. 743-800, 2021. doi:10.3389/fphys.2021.743800
- [2] R. A. Dewanti and I. Hermawan, "A comparative study on measuring running speed using timing gates and stopwatches to reduce human error," *Journal of Physical Education and Sport*, vol. 22, no. 12, pp. 3080-3085, 2022. doi:10.7752/jpes.2022.12390
- [3] P. Mamo, "Timing Gates vs. Hand-Held Stopwatches: The Superiority in Measuring Sprint Performance," 2023. [Online]. Available: <https://petermamo.com/blog/timing-gates-vs-hand-held-stopwatches-the-superiority-in-measuring-sprint-performance/>. [Accessed 3 April 2025].
- [4] A. Moreno-Azze, D. López-Plaza, F. Alacid and D. Falcón-Miguel, "Validity and Reliability of an iOS Mobile Application for Measuring Change of Direction Across Health, Performance, and School Sports Contexts," *Applied Sciences*, vol. 15, no. 4, p. 1891, 2025. doi:10.3390/app15041891
- [5] H. Uysal, A. Ojeda-Aravena, M. Ulaş, E. Báez-San Martín and R. Ramirez-Campillo, "Validity, Reliability, and Sensitivity of Mobile Applications to Assess Change of Direction Speed," *Journal of Human Kinetics*, vol. 88, pp. 217-228, 2023. doi:10.5114/jhk/167465
- [6] H. S. Krishan, J. J. Mathunny, K. Govindasamy, A. Devaraj and V. Karthik, "Inter-device reliability of photo finish: Android based smartphone application for the measurement of running speed," in *AIP Conference Proceedings*, Chennai, India, 2023. doi:10.1063/5.0126231
- [7] R. K. Thapa, B. Sarmah, T. Singh, G. S. Kushwah, Z. Akyildiz and R. Ramirez-Campillo, "Test-retest reliability and comparison of single- and dual-beam photocell timing system with video-based applications to measure linear and change of direction sprint times," *Proceedings of the Institution of Mechanical Engineers, Part P*, 2023. doi:10.1177/17543371231203440
- [8] F. J. Barrera-Dominguez, P. A. Jones, B. J. Almagro and J. Molina-Lopez, "Validity and Inter-Device Reliability of an Artificial Intelligence App for Real-Time Assessment of 505 Change of Direction Tests," *European Journal of Sport Science*, vol. 25, no. 2, p. e12252, 2025. doi:10.1002/ejsc.12252
- [9] S. Altmann, L. Ruf, M. Backfisch, M. Thron, A. Woll, L. Walter, D. Kaul, L. Bergdolt and S. Härtel, "Assessing maximal sprinting speed in soccer – criterion validity of commonly used devices," *Science and Medicine in Football*, pp. 1-6, 2024. doi:10.1080/24733938.2024.2441321



- [10] B. Zajac, P. Gaj and J. Zięba, "Concurrent Validity and Inter-Rater Reliability of Hand-Held Measurements of Maximal Sprint Speed," *Journal of Kinesiology and Exercise Sciences*, vol. 32, no. 100, pp. 15-21, 2022. doi:10.5604/01.3001.0016.1225
- [11] A. Skujytė, I. Lukonaitienė, J. Stanislovaitienė, V. Šilinskas, K. Bradauskienė, G. Mamkus and S. Kamandulis, "Reliability and validity of sprint performance using the Alex7 motorized device," *Frontiers in Sports and Active Living*, vol. 6, 2024. doi:10.3389/fspor.2024.1412861
- [12] R. van den Tillaar, M. Haugen and H. Falch, "A Comparison of Sprint Mechanical Parameters Measured With Timing Gates and a Laser Gun," *Frontiers in Sports and Active Living*, vol. 4, 2022. doi:10.3389/fspor.2022.877482
- [13] L. Marco-Contreras, A. Bataller-Cervero, H. Gutiérrez, J. Sánchez-Sabaté and C. Berzosa, "Analysis of the Validity and Reliability of the Photo Finish® Smartphone App to Measure Sprint Time," *Sensors*, vol. 24, no. 20, p. 6719, 2024. doi: 10.3390/s24206719



THE PRESENCE OF TECH COMPANIES IN SUPER BOWL ADS FROM 2021 TO 2025

Aleksandra Belačić*,
[0009-0003-1152-1645]

Slavko Alčaković
[0000-0003-4062-7214]

Singidunum University,
Belgrade, Serbia

Abstract:

The main idea behind this paper is to examine the presence of companies belonging to the tech industry at the Super Bowl, the annual championship game of the U.S. National Football League, which enjoys the status of a mass cultural event both in its home country and all across the globe. In the theoretical part of the paper, the concept of the Super Bowl is presented alongside statistical data that highlight its global significance and, by extension, the prestige associated with advertising during its commercial breaks. Additionally, this part of the paper outlines the history of Super Bowl advertising within the broader tech sector, focusing on IT companies. The paper's practical part presents the results of research that involved reviewing all ads aired during the Super Bowl over the past five years – 312 in total – with the aim of determining the representation of companies belonging to the tech sector. The results are presented in table form, followed by an explanation of the figures and additional observations on the strategies tech companies use with the aim of successfully appealing to the Super Bowl audience.

Keywords:

Super Bowl, Advertising, Television Commercials, Tech Industry, Brand Promotion.

INTRODUCTION

The advertising landscape was transformed to a great extent after the emergence of social networks, which provided opportunities for targeting people based on their location or demographic parameters like their gender, interests, or even job title. [1] Apart from that, it turned out that advertising via the Internet is on a whole other level of cost-effectiveness than traditional one, and that it also offers other cutting-edge functions, completely unheard of in the era of classic television, the most important one being measurability, that is to say, the ability to get detailed analytics for every advert. [2]

The current circumstances in the advertising industry can be described by the term "attention economy", coined to refer to the fact that companies nowadays must compete for consumer attention. [3] This practically means that there has been a reversal of the old and obviously outdated concept of "information economy", i.e., adaptation to circumstances in which information is plentiful and attention is a limited and therefore scarce resource. [4]

Correspondence:

Aleksandra Belačić

e-mail:

aleksandra.belacic.22@singimail.rs





So, at present, information is cheap and attention is expensive – and it's up to advertisers to find the best way to stand out, grab attention of the public and encourage them to try the offered products and services. It is a matter of debate whether television advertising is still a proper way of doing so, with one line of thought suggesting it's not, and that it often makes a much smaller impact on sales than it was previously estimated. [5] According to the proponents of said theory, in many cases, companies spent more on commercials than they earned back in extra sales, resulting in a negative return on investment. [5] According to the second theory, it doesn't matter where the advertising takes place as people remember brands, they've seen in Internet ads just as well as those seen in TV ads. [6] The third theory suggests that, in spite of the growth of new technologies and digital media, television still enjoys the status of the most effective advertising platform. [7]

Although television commercials are increasingly being replaced by those on the Internet, the Super Bowl, the single most prestigious event for traditional advertising, remains unshaken by this burgeoning trend. Since its establishment 58 years ago, the spectacular broadcast of the finals of the American NFL league has had consistently colossal viewership among U.S. viewers. [8] According to the author Lotz, the fact that the Super Bowl still has the power to gather the nation around the same program, as in the golden era of television, makes it as much a high-profile advertising event as it is a championship game. [9]

Huge viewer base, along with the lasting buzz that all commercials aired during each and every Super Bowl bring with them, motivates the advertising industry's top dogs to pay fortunes for mere seconds of advertising space. In addition, they skilfully employ humour or sentimentality during the process of creating their ads, all in an attempt to make long-lasting impact by engaging with viewers' emotions. [10]

The idea behind this paper is to examine the representation of companies from the tech sphere (that is, those operating in areas of start-ups, innovations, electronics, gadgets, applications, artificial intelligence, robotics, etc.) at the Super Bowl in the span of the last five years, as well as to further investigate the strategies they apply in advertising at such a highly visible global showstopper.

2. THE IMPORTANCE OF THE SUPER BOWL

Founded back in 1967 [11], the Super Bowl, the NFL season finale, where the two best teams compete for the championship title, is always held at the peak of the winter season [12], while the location of the event is determined based on the results of the bidding process, where the hosting city that meets specific standards and criteria is chosen from among the candidates. [13]

With a live audience that constantly varies between 60 and 70k viewers (not counting the year 2021 when it was significantly reduced due to the COVID-19 pandemic) [14] and an organizing budget being in between the staggering amounts of 50 and 60 million US dollars [15], the Super Bowl is classified as a sports mega-event, which not only manages to draw massive crowds but also to attract extensive attention. [16] The results of said attention include a record-breaking TV viewership that reached approximately 127.7 million during Super Bowl 2025 [17], as well as a staggering 2.83 billion engagements related to the same event, made across the most popular social networks such as Instagram, X, and YouTube. [18]

From the TV viewership-related data collected over the past 35 years, it can be seen that the numbers, although they have not always followed a growing path, have remained consistently high [17], which is not surprising since the Super Bowl has the status of the most viewed televised event in the United States. [19] Popular across all 50 states, the Super Bowl also displays the effect called "home-city" or "host-city," which practically means that the viewership of the game is roughly 20% higher than normal in the city where it is held. [20] Also, despite the fact that American football is a U.S. national sport that is rarely played abroad, the Super Bowl enjoys surprising popularity all over the world, even in territories where this sporting activity is non-existent. How much of a social phenomenon the Super Bowl is, is best illustrated by data from 2022, when its global viewers number was estimated to surpass 208 million, with 130 countries having purchased the streaming rights. [21]

The fact that the Super Bowl viewership, despite slight fluctuations, consistently stays at an extraordinarily high level allows the price of an advertising second to only ever go up. At the time of the first Super Bowl, a 30-second commercial spot cost between \$37,500 and \$42,500 [22], and in 2025 it skyrocketed to a staggering \$7 million, while at least ten of the commercial spots were later resold for over \$8 million each. [23]



According to the latest data, Super Bowl viewers are predominantly male (57%) and Caucasian (68%) with an average age of 49, median household income of \$74K (on par with the national average), and solid financial footing (35% of them surpass the \$250,000 net worth threshold). [24] This exceptional economic influence quite logically motivates companies to use the platform of the Super Bowl to advertise products that specifically satisfy said demographic.

Despite the fact that the cost of airing Super Bowl commercials has increased drastically over the years, advertisers are still motivated to spend millions of dollars creating and producing high-quality ads [25], a trend that has turned the game itself into only part of the attraction.

3. ADVERTISING AT THE SUPER BOWL

Each year, long before it is known which two teams will face off in the finals [20], advertisers go into a total frenzy to secure Super Bowl commercial slots at seemingly sky-high prices. [25] For example, the demand for ad slots in 2010 was so intense that 80% of them were sold out eight months prior to the big game. [26] Exorbitant advertising price also acts as an eliminator, allowing only top-ranking companies to appear at this prestigious event.

Prominent companies are interested in advertising during the Super Bowl not only due to the great exposure it offers, but also because viewers really pay close attention to the ads broadcast during this event, often gathering in front of the screen once they begin, which is the opposite of the typical behaviour when people tend to focus on other things or even leave the room during the commercial blocks. [26] It is precisely this characteristic that gives the Super Bowl the ability to greatly influence how consumers perceive brands, even those that are already well-established. [26]

In practice, although the Super Bowl follows the same rules as other NFL games, its broadcasts are more carefully directed due to the larger number of sponsors, longer duration and higher cost of ads, as well as the greater interest and focus of the audience. Therefore, in 2025, the organizers decided to add extra breaks to the standard number of 18 advertising blocks, bringing the total to 20. [27]

The commercials aired at the Super Bowl receive evaluation both from the general public and by a focus group formed by USA Today's Ad Meter platform. Publishing the focus group's results, which rank ads by likeability, helps creators analyse the elements of the best ads and gain insight into what works with the public.

4. THE HISTORY OF TECH AND IT COMPANIES' PRESENCE AT THE SUPER BOWL

The Super Bowl commercial breaks were traditionally filled with ads for food, alcoholic beverages, and automotive brands. [28] The first significant change in that established concept occurred during Super Bowl 2000, which was so packed with internet-related companies that it was dubbed the "dot-com Super Bowl" [29] or "e-bowl." [30] This happened as a logical outcome of the dot-com bubble, which began in 1995, and was fuelled by the already existing interest of businesses and entrepreneurs in capitalizing on the rapid growth of the Internet. [31] At the aforementioned Super Bowl, where 30-second advertising slots cost an average of \$2.2 million each, there were 36 advertisers, and 17 of them (47.22%) were internet-operated companies. [32]

Advertisements of IT companies, which are part of the larger tech industry, were present at the Super Bowl even before 2000, but to a significantly lesser extent – three in 1999, one in 1998, and one in 1997. [30] Practically, the end of the 20th century was the Internet land-grab era that led dot-coms to pursue extravagant advertising campaigns. [33] Many of those newly established businesses quickly jumped at the chance to advertise during the Super Bowl, hoping that such massive exposure would catapult them to levels of expertise and influence typical of Fortune 100 companies. [34]

The results of the dot-com companies' broke-or-break media strategy were, at best, mixed. One research study pointed out that lesser-known companies, relatively new to both the marketplace and Super Bowl advertising, had the greatest positive change in brand familiarity. At the same time, it was found that Super Bowl exposure brings little gain in brand familiarity or likeability to already well-known brands. [34] In other words, ads placed in commercial breaks of the Super Bowl can raise a company's stock price, but it typically happens only when the brand is not well known or has been suffering from a damaged reputation. [35]

The example of pets.com, an internet company that advertised at the Super Bowl 2000 only to collapse nine months later, made it clear that the marketing approach of the dot-com companies was more quantity-over-quality based and that the results would have been much better if the money had been spent more effectively on ads through targeted media. [36] Consequently, dot-com marketing is often recalled as a true overspending bacchanalia. [37]



In 2000, the U.S. TV network ABC estimated the Super Bowl audience at 130.7 million viewers, and while these numbers guaranteed high visibility for advertisers, the risks involved were enormous. [32] In other words, Super Bowl advertising was at that time a game of big risk and big reward that could put some companies into the spotlight, but also bring them crashing down. [38] The era when money was being recklessly spent therefore brought one very important lesson: those who want to advertise at the Super Bowl have to have enough insurance to allow for failure. [38]

5. RESEARCH AND RESULTS

The central idea behind our research was to evaluate the presence of tech companies at the Super Bowl in the last five years, that is, to determine whether the trend is rising, declining, or steady, as well as to determine which companies are advertising at this event and what their campaign tactics are. The initial idea was to collect the data needed for our analysis from the site Ad Meter, known for years as a reference database where all ads broadcast during the Super Bowl were located, along with their likeability score. However, due to policy changes on this site, now only the best-rated ads are displayed, instead of the entire base, which is why, for the purposes of this research, we opted to use the Super Bowl Ads website (<https://www.superbowl-ads.com/>), which contains the entire base of advertisements played during Super Bowl commercial breaks starting from 1998.

By inspecting the aforementioned database, we determined that in the period from 2021 to 2025, a total of 312 ads were broadcast during Super Bowl breaks, and we then went on assessing them in detail with the aim of calculating how many of them belong to the tech sector. It is important to point out that drawing a sharp line between what is and what is not a tech ad or industry is a very challenging task. For the purposes of this research, we have taken the position that, apart from core tech

companies, broader digital companies which operate mostly via online platforms, as well as digital commerce websites and social networks, can also be classified in that category. Having learned that other sources reporting on tech ads also include media companies primarily focused on tech-powered content production and distribution – such as Netflix, Paramount+, Peacock, etc. – we decided to consider them as well.

The findings of our research are presented in Table 1, which displays the total number of ads aired during each edition of the Super Bowl between 2021 and 2025, as well as the total number and percentage of technology ads broadcast during each of those years. From those figures, it can be seen that the ads belonging to the technology sector consistently made up more than 30% of all Super Bowl ads, reaching their peak figure of 36.23% in 2022. During the five-year period covered by our research, a total of 312 advertisements were aired during Super Bowl commercial breaks, and 104 of them were related to technology, which is why it can be concluded that the tech industry has joined food, beverage, and automotive industries in becoming one of the leading Super Bowl advertisers. By analysing the ads, we also determined that the structure of tech advertisers is extremely diverse, ranging from powerhouses like Google and YouTube to core tech companies such as OpenAI, and extending to predominantly internet-enabled businesses and e-commerce platforms like Booking, Temu, and Amazon, telecom companies like T-Mobile and Verizon, and video streaming services such as Netflix.

Strategies used by tech companies to make their products and services appeal to the public may differ, but one strong trend emerges among all of them, that being the growing focus on artificial intelligence. The strongest brands from the technological sphere have begun to design their ads in a way that highlights the possibility of practical application of their achievements in the AI domain. Most prominent examples include Microsoft's 2024 ad for its AI-powered tool, Copilot, that helps its users achieve learning and business tasks, Meta's 2025

Table 1. Overview of the Presence of Technology Companies at the Super Bowl (2021–2025)

Year	Total number of ads	Number of technology ads	Percentage of technology ads
2021	55	18	32.73%
2022	69	25	36.23%
2023	62	20	32.26%
2024	60	20	33.33%
2025	66	21	31.82%



ad for AI-powered Ray-Ban smart glasses capable of providing real-time information about viewed objects, Google's 2025 ad emphasizing the everyday utility of its AI assistant Gemini Live, as shown via a job interview preparation scenario, and OpenAI's 2025 ad presenting ChatGPT's capabilities in assisting with writing-related tasks.

6. CONCLUSION

Our research, which relates to the period from 2021 to 2025, shows that during that time frame, the ads of tech companies accounted for a significant portion of the advertising space at the Super Bowl, with more than 30% participation each year. These results coincide with the wider social picture, which is characterized by an exponential growth in the integration of new AI-based technologies into everyday household and work-related tasks. The obtained results show that tech companies, along with traditional advertisers, became regular participants in the commercial segment of this prestigious sporting mega event, thereby confirming their financial power and their desire to move away from the boundaries of technologically specialized audiences and approach the mass population.

The performance of technology companies is characterized by diversity in terms of the structure of the organizations themselves, but also in terms of the strategies they use, with majority of them relying on artificial intelligence in order to demonstrate that their exceptionally high-tech products and services can be integrated into everyday life.

It is important to note that this research has certain limitations that do not affect its results to a greater extent, concerning the criteria according to which the authors defined what tech companies are. Likewise, the limiting factor may be the quality of the used reference database, that is, whether it really includes all broadcast advertisements. A recommendation for further research on this topic would be to repeat the research in a few years to include new trends, as well as to take a look at the business results of tech companies that advertised on the Super Bowl in order to see whether they truly benefited from that exposure.

REFERENCES

- [1] H. Albadri, "The Convergence of Traditional Media to the Digital Communicative Environment- The Reality and Gap," *Information Sciences Letters*, vol. 12, no. 4, pp. 1827-1839, 2023.
- [2] G. Köse, "An Innovative Approach Within the Digital Marketing Spectrum: Programmatic Advertising as an Output of Artificial Intelligence," in *International Studies and Evaluations in the Field of Communication Sciences*, Ankara, Turkey, Serüven Yayinevi, 2024, p. 50.
- [3] E. C. Parisi and F. Parisi, "Rethinking Remedies for the Attention Economy," SSRN, 2021.
- [4] P. Crogan and S. Kinsley, "Paying Attention: Toward a Critique of the Attention Economy," *Culture Machine*, vol. 13, p. 1-29, 2012.
- [5] B. Shapiro, G. Hitsch and A. Tuchman, "TV Advertising Is Usually Not Worth It," Kellogg Insight, 13 2021. [Online]. Available: <https://insight.kellogg.northwestern.edu/article/tv-advertising-is-usually-not-worth-it>. [Accessed 14 2025].
- [6] M. Draganska, W. R. Hartmann and G. Stanglein, "Internet vs. TV Advertising: A Brand Building Comparison," *Journal of Marketing Research*, vol. 51, no. 5, p. 578-590, 2014.
- [7] F. Findley, K. Johnson, D. Crang and D. W. Stewart, "Effectiveness and Efficiency of TV's Brand-Building Power: A Historical Review," *Journal of Advertising Research*, vol. 60, no. 2, p. 131-149, 2020.
- [8] Sports Media Watch, "Super Bowl Ratings History (1967-present)," Sports Media Watch, 4 2017. [Online]. Available: <https://www.sportsmediawatch.com/super-bowl-ratings-historical-viewership-chart-cbs-nbc-fox-abc/>. [Accessed 14 2025].
- [9] A. D. Lotz, *The Television Will Be Revolutionized*, Second Edition ed., New York: New York University Press, 2014.
- [10] P. Beede, "The Growing Creative Challenge for Advertising Engagement: An Analysis of Super Bowl Ad Meter Likeability Scores from 2013 to 2024," *Journal of Marketing Communications*.
- [11] Q. Feng, "Analyzing the Economic Impact of Super Bowls: A qualitative framework for holding mega sports events in Asia," in *Proceedings of the 9th International Conference on Financial Innovation and Economic Development (ICFIED 2024)*, 2024.
- [12] J. Cuneen and J. S. Fink, "Marketing Motown: Detroit Sold Cold, and Super Bowl XL Was a Winter Blast," *Sport Marketing Quarterly*, vol. 15, no. 4, 2006.
- [13] A. D. Smith and A. A. Smith, "Exploring the Service Location Strategies Behind Super Bowl Venue Selection," *Services Marketing Quarterly*, vol. 29, no. 4, p. 66-84, 2008.



- [14] Statista Research Department, "Super Bowl attendance in the U.S. 1967-2025," Statista, [Online]. Available: <https://www.statista.com/statistics/1362951/super-bowl-attendance/>. [Accessed 2025 4 1].
- [15] F. Monthly, "How Much Does the Super Bowl Cost to Run?," Finance Monthly, 10 2 2025. [Online]. Available: <https://www.finance-monthly.com/2025/02/how-much-does-the-super-bowl-cost-to-run/>. [Accessed 1 4 2025].
- [16] E. Mangeloja, "Economic Impacts of Sports Mega-Events: Introduction to Economic Analysis of Sports Mega-Events," in *Gender, Politics and Evolution of Sports: Nordic-Baltic Perspectives on History of Sports*, T. Roiko-Jokela and H. Roiko-Jokela, Eds., Jyväskylä, Finland: Finnish Society for Sport History, 2024, pp. 16-35.
- [17] J. Stoll, "Super Bowl TV viewership in the U.S. 1990-2025," Statista, 13 2 2025. [Online]. Available: <https://www.statista.com/statistics/216526/super-bowl-us-tv-viewership/>. [Accessed 1 4 2025].
- [18] S. Karlovitch, "Super Bowl LIX: Analyzing the Game's Advertising and Engagement Data," Marketing Dive, 10 2 2025. [Online]. Available: <https://www.marketingdive.com/news/super-bowl-lix-advertisement-engagement-data/732443/>. [Accessed 1 4 2025].
- [19] P. Minnium, "Super Bowl Ad Rankings: Trust Your Gut or Employ Rational Thinking?," MarTech, 26 2 2018. [Online]. Available: <https://martech.org/super-bowl-ad-rankings-trust-gut/>. [Accessed 1 4 2025].
- [20] S. Stephens-Davidowitz, H. Varian and M. D. Smith, "Super Returns to Super Bowl Ads?," *Quantitative Marketing and Economics*, vol. 15, no. 1, p. 1-28, 2017.
- [21] S. Alčaković, "Super Bowl Ads in the Period from 2017-2022," *The European Journal of Applied Economics*, vol. 20, no. 1, pp. 39-51, 2023.
- [22] E. McDowell and R. Su, "How Super Bowl Ad Costs Have Skyrocketed Over the Years," Business Insider, 2 2 2020. [Online]. Available: <https://www.businessinsider.com/super-bowl-ad-price-cost-2017-2>. [Accessed 1 4 2025].
- [23] D. Hayes, "Fox Breaks \$8M Barrier For A 30-Second Super Bowl Ad, Selling At Least 10 Spots At That Level," Deadline, 29 1 2025. [Online]. Available: <https://deadline.com/2025/01/fox-super-bowl-commercials-8-million-30-seconds-1236270747/>. [Accessed 1 4 2025].
- [24] M. Culture, "Super Bowl Enthusiasts: Audience Insights Report," Media Culture, 3 2 2025. [Online]. Available: <https://www.mediaculture.com/insights/audience-insights-report-super-bowl-enthusiasts>. [Accessed 1 4 2025].
- [25] R. Yelkur, C. Tomkovick, A. Hofer and D. Rozumalski, "Super Bowl Ad Likeability: Enduring and Emerging Predictors," *Journal of Marketing Communications*, vol. 19, no. 1, pp. 58-80, 2011.
- [26] W. R. Hartmann and D. Klapper, "Super Bowl Ads," 2016.
- [27] K. Belson, "During N.F.L. Games, Going to Commercial Requires Its Own Playbook," The New York Times, 8 2 2025. [Online]. Available: <https://www.nytimes.com/2025/02/08/business/nfl-games-commercials-super-bowl.html>. [Accessed 1 4 2025].
- [28] S. Lingeswaran, "Explainer: The Evolution of Super Bowl Ads," Sportcal, 7 2 2023. [Online]. Available: <https://www.sportcal.com/features/explainer-the-evolution-of-super-bowl-ads/>. [Accessed 1 4 2025].
- [29] L. K. Fuller, *Celebrating the Super Bowl: Programs, Profits, Parties*, Routledge, 2022.
- [30] S. Elliott, "The Media Business: Advertising; Move Over, Dot-Coms. There Are Other First-Time Advertisers Gearing Up for the Super Bowl," The New York Times, 18 1 2000. [Online]. Available: <https://www.nytimes.com/2000/01/18/business/media-business-advertising-move-over-dot-coms-there-are-other-first-time.html>. [Accessed 1 4 2025].
- [31] R. R. Panko, "IT Employment Prospects: Beyond the Dotcom Bubble," *European Journal of Information Systems*, vol. 17, p. 182-197, 2008.
- [32] S. Elliott, "Shocking Defeats and Other Super Bowl XXXIV Marketing Memories," The New York Times, 1 2 2000. [Online]. Available: <https://www.nytimes.com/2000/02/01/business/advertising-shocking-defeats-and-other-super-bowl-xxxiv-marketing-memories.html>. [Accessed 1 4 2025].
- [33] D. Laffey, "The Rise and Fall of the Dot Com Entrepreneurs," Canterbury Business School, University of Kent, Canterbury, 2004.
- [34] M. G. Russell, J. A. Fortunato, L. Valencia and N. M. Burns, "Brand Perceptions of TV Commercials During Super Bowl XXXVIII," *Clickin Research*, Inc.; University of Texas at Austin.
- [35] M. Palmquist, "Doing the Super Bowl Ad Math," Strategy+Business, 1 2 2017. [Online]. Available: <https://www.strategy-business.com/blog/Doing-the-Super-Bowl-Ad-Math>. [Accessed 1 4 2025].
- [36] J. Thornton and S. Marche, "Sorting Through the Dot Bomb Rubble: How Did the High-Profile E-Tailers Fail?," *International Journal of Information Management*, vol. 23, no. 2, p. 121-138, 2003.
- [37] E. Anderson, *Social Media Marketing: Game Theory and the Emergence of Collaboration*, Springer-Verlag Berlin Heidelberg, 2010.
- [38] A. Steinbach, "Advertising in the Super Bowl: Worth the Cost?."



STUDENT SESSION



STUDENT SESSION

UTILIZING BUSINESS ANALYTICS, BIG DATA, AND VISUALIZATION FOR SALES PERFORMANCE OPTIMIZATION

Aleksandar Radivojević*,
[0009-0008-0573-2362]

Miloš Mravik,
[0000-0001-5442-3998]

Marko Šarac
[0000-0001-8241-2778]

Singidunum University,
Belgrade, Serbia

Abstract:

This research paper examines the role of business analytics and reporting in the sales sector, utilizing big data. The main goal is to analyse sales data, identify key insights, and present the results through visualization using Tableau.

The introductory section discusses the research subject, objectives, and challenges that may arise when applying business analytics in sales. Types of content explores the technologies used for handling big data, which are crucial for the efficient storage, processing, and analysis of large datasets. These technologies include distributed systems like Hadoop, data warehouses such as Red Lake, and tools for data analysis and integration, including SQL Server, SSIS, and SSAS. The section also highlights the role of data visualization tools like Tableau and Power BI in presenting key insights and supporting business decision-making. The methodology describes the steps for collecting and analysing sales data, along with an overview of data visualization tools such as Tableau. A dedicated section provides a detailed description of the database, including its structure, and data types.

The data analysis covers various types of sales performance, customer, and product analyses. Data visualization and interpretation are presented through Tableau, focusing on result interpretation and their application in business decision-making. The thesis concludes with a summary of key findings.

Keywords:

Business Analytics, Business Reporting, Tableau, Data Visualization, Distributed Systems, Big Data.

INTRODUCTION

The subject of this research is the application of business analytics and reporting, working with big data, all within the context of the sales sector. The focus is on the analysis and visualization of sales data and improving the presentation of the final solution in the form of data products. Today, business analytics is achieved through two main groups of tools: business intelligence tools and big data analytics tools. Business Intelligence (BI) utilizes techniques such as data mining, data visualization, and specialized tools to enable organizations to make informed decisions. Big Data analytics tools allow organizations to collect, store, process, and analyse large amounts of data to gain valuable insights. These tools help extract valuable information from complex and unprocessed data.

Correspondence:

Aleksandar Radivojević

e-mail:

aleksandar.radivojevic.20@singimail.rs





The aim is to explore and demonstrate how business analytics and reporting can significantly impact the analysis of sales data and improve sales strategies. Specifically, it will focus on the use of data visualization tools such as Tableau. It includes identifying key patterns and insights from the data that can contribute to a better understanding of sales trends, support informed decision-making, and enhance sales strategies.

2. TYPES OF CONTENT

2.1. BUSINESS ANALYTICS

Business analytics is the practice of analysing and interpreting data to gain insights into business performance and identify areas for improvement. This process can be enhanced with the use of various techniques and tools, such as predictive analytics, machine learning, and natural language processing, which empower analysts and business users to glean insights from structured and unstructured data sources. [1]

The importance of business analytics lies in the fact that it is crucial for every organization to have informed decisions to improve performance. When businesses have insights into their data, they can automatically optimize their processes to reduce costs, enhance customer experiences, identify new growth opportunities, and even predict future trends and anticipate customer needs.

2.2. BUSINESS REPORTING

Business reporting refers to the practice of gathering, analysing, and presenting data related to organizational activity and performance. In today's business world, where competition is rampant, the role of business reporting cannot be overemphasized. It allows for informed decision-making by providing accurate information that is critical to the success of any corporation.

Through reporting, companies can easily track their performance, identify trends, and make data-driven decisions that can impact organization-wide operations. It is essential, therefore, that organizations allocate sufficient resources to this process to ensure that they have access to real-time insights that enable them to stay competitive in the dynamic business environment. [2]

2.3. BIG DATA AND DISTRIBUTED SYSTEMS

With the rapid growth of Internet users, data generation has increased exponentially. This data comes from millions of messages sent and received via platforms like WhatsApp, Facebook, and Twitter, as well as trillions of photos taken and countless hours of videos uploaded to YouTube every minute. According to a recent survey, 2.5 quintillion (2,500,000,000,000,000,000 or 2.5×10^{18}) bytes of data are generated daily. This massive volume of data is referred to as "big data." However, big data is not just about size—it is a broad term for data that is vast, complex, and can be structured or unstructured, arriving at high velocity. Notably, 80% of today's data has been generated in just the past few years. The continuous expansion of big data is driven by the ever-increasing volume of produced information and the growing need to capture and analyse it. [3]

The defining characteristics of Big Data are commonly represented by the 5V model:

1. Volume – the vast amount of data.
2. Variety – the diverse types of data.
3. Velocity – the speed at which data is generated and processed.
4. Veracity – the trustworthiness and quality of data.
5. Value – the usefulness of the data for decision-making and insights.

Data Warehouses: Red Lake is a data warehouse (Data Warehouse) used for processing and analysing large volumes of data (Big Data). Although it is not as widely known as traditional solutions like Snowflake, Google Big Query, or Amazon Redshift, Red Lake is increasingly utilized in DataOps and analytical environments where fast data storage and real-time data processing are required. It is designed to handle the challenges of Big Data by providing a platform that can scale quickly and support high-speed processing, essential for modern data-driven applications.

Distributed Systems: Apache Pig is a distributed system for processing large datasets, used for data analysis and transformation within the Hadoop ecosystem. Developed by Yahoo!, Pig enables users to write complex data processing tasks using a simple scripting language called Pig Latin. This language is designed to abstract the complexity of traditional MapReduce programming, making it easier for developers to process large datasets. How Apache Pig Works:



Data Analysis and Integration Tools: SQL Server, SSIS (SQL Server Integration Services), and SSAS (SQL Server Analysis Services) are tools that enable data integration, analysis, and transformation. SQL Server – A relational database developed by Microsoft. It enables data storage, management, and querying using the T-SQL language. SSIS (SQL Server Integration Services) – A tool for ETL processes (Extract, Transform, Load). It is used for data integration, cleaning, and migration from various sources. SSAS (SQL Server Analysis Services) – A service for analytical data processing. It enables the construction of OLAP cubes and data models, which facilitates advanced analytics and BI reporting.

Visualization Tools: Tableau and Power BI are the most popular data visualization tools that enable users to create visual representations of data easily. These tools simplify the analysis of large datasets and allow for informed decision-making through interactive visualizations. Tableau – A data visualization tool that enables interactive and intuitive charts, dashboards, and analytics. It is suitable for data exploration and storytelling. Power BI – A Microsoft BI tool for reporting and analytics, allowing connection to various data sources, creation of interactive dashboards, application of DAX expressions, and integration with SQL Server and other Microsoft tools. [4]

2.4. STRUCTURE OF DATABASE

The database used for analysis and reporting is orders_frostonline.xlsx. It uses the .xlsx format, which stands for Excel Spreadsheet in XML format and was first introduced in Excel 2007 as a replacement for the older .xls format. .xlsx files are compressed ZIP archives containing multiple XML files and other types of data organized in a structure that represents worksheets, tables, charts, formulas, and other Excel functionalities.

It contains eight tables: Categories, Customers, Employees, Orders, Orders details (details of each individual order), Products, Shippers, and Suppliers. These tables are interconnected through primary keys. When we want to link two or more tables in databases, they need to have a key, meaning a column or set of columns that will allow the tables to be connected. The primary key is a unique identifier within a table. Each row in that table has a unique value in the column representing the primary key.

Table 1 shows information about Table Categories:

Table 1. Table Categories

CategoryID	CategoryName	DescriptionText
1	Beverages	Soft drinks, coffees, teas, beers, and ales
2	Condiments	Sweet and savoury sauces, relishes, spreads, and seasoning
3	Confections	Desserts, candies, and sweet breads
4	Dairy Products	Cheeses

The table below, Table 2, shows information regarding Table Customers:

Table 2. Table Categories

Customer ID	Customer-Name	ContactName	Adress	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
3	Berglunds snabbkop	Christina Berglund	Bergusvagen 8	Lulea	S-958 22	Sweden



Table 3 is showing information regarding Table Employees:

Table 3. Table Employees

EmployeeID	LastName	FirstName	BirthDate	Photo	Notes
1	Davolio	Nancy	1968-12-08	EmplD1.pic	Education includes...
2	Fuller	Andrew	1952-02-19	EmplD2.pic	Andrew received...
3	Leverling	Janet	1963-08-30	EmplD3.pic	Janet has...

In Table 4, we can see information about Table Orders:

Table 4. Table Orders

OrderID	CustomerID	EmployeeID	OrderDate	ShipperID
10248	90	5	1996-07-04	3
10249	81	6	1996-08-04	1
10250	34	4	1996-09-08	2
10251	84	3	1996-10-08	1

Table 5 is showing information regarding Table OrdersDetails:

Table 5. Table OrdersDetails

OrderDetail ID	OrderID	ProductID	Quantity
1	10248	11	12
2	10248	42	10
3	10248	72	5
4	10249	14	9

In Table 6, we can see information about Table Products:

Table 6. Table Products

ProductID	ProductName	SupplierID	CategoryID	Unit	Price
1	Chais	1	1	10 boxes x 20 bags	18
2	Chang	1	1	24 – 12 oz bottles	19
3	Aniseed Syrup	1	2	12 – 550 ml bottles	10
4	Chef Antons Cajun Seasoning	2	2	48 – 6 oz jars	22

Table 7 and Table 8 are showing information regarding Table Shippers and Table Suppliers:

Table 7. Table Shippers

ShipperID	ShipperName	Phone
1	Speedy Excpres	(503) 555-9831
2	United Package	(503) 555-3199
3	Federal Shipping	(503) 555-9931



Table 8. Table Suppliers

SupplierID	SupplierName	ContactName	Adress	City	Postal-Code	Country	Phone
1	Exotic Liquid	Charlotte Cooper	49 Gilber St.	London	EC1 4SD	UK	(171) 555-2222
2	New Orleans Cajun Delights	Shelley Burke	P.O. Box 78934	New Orleans	70117	USA	(100) 555-4822
3	Grandma Kellys Homestead	Regina Murphy	707 Exford Rd.	Ann Arbor	48104	USA	(313) 555-5735

2.5. WORK IN TABLEAU

In this research, the connection and integration of sales-related datasets enable comprehensive analysis and visualization. By utilizing Tableau, the data from multiple tables is combined, processed, and presented through interactive dashboards. This tool allows for seamless data exploration, facilitating the identification of key sales trends, performance metrics, and correlations. The ability to create dynamic visual representations enhances data-driven decision-making, making Tableau an essential component of modern business analytics and reporting. [5]

2.5.1. The physical layer of connection

For everything to function properly, each of these tables must be physically connected through common keys via inner joins. The Customers table is connected to the Orders table through the CustomerID field, which is present in both tables; the Orders table is connected to the Employees table through EmployeeID; to the Order Details table via OrderID; and to the Shippers table through ShipperID. The Order Details table is connected to the Products table via the ProductID field, while the Products table is connected to the Categories and Suppliers tables through the CategoryID and SupplierID keys, respectively.

2.5.2. Dashboards

By combining dimensions and measures in Tableau, you can create meaningful visualizations:

- Rows often contain “Measures”, displaying quantities (e.g., total sales).
- Columns often contain “Dimensions”, categorizing data (e.g., months or regions).

For example, if you want to see sales by month, the Month (Dimension) would be placed in Columns, while Sales (Measure) would be in Rows. This creates a chart where the X-axis represents months, and the Y-axis represents sales figures. Some measures in Tableau are automatically generated, while others need to be created using SQL queries, depending on what you want to display. Figure 1 shows a chart named “Customers in Country”, representing the total number of customers in each country. Visually, it is displayed as a world map, where each country with customers is shaded in dark brown or lighter brown, depending on the number of customers. Countries without customers are shown in grey colour. When hovering over a country shaded in brown, a small tooltip appears, displaying the country's name and the total number of customers from that country.

Figure 2 shows the table “Customer Spent” that displays key information about customers related to their spending. It includes the following details: Customer ID, Full name, Residential address, and total amount spent by the customer. Different colours in the Sum (Price) column visually highlight customer spending – a darker colour indicates a higher spending amount. This visualization allows for the quick identification of high-spending customers, making it easier to target marketing and sales efforts more effectively.

In Figures 3 and 4, different data analysis techniques are illustrated. Figure 3 presents forecasting, which is useful for predicting trends in sales, revenue, inventory, weather conditions, and other areas where planning ahead is essential. In this study, forecasting was used to predict the number of customers from 1999 to 2002 based on data from 1997 to 1999. This means that a longer historical period is used to make predictions for an upcoming period. The curve illustrates the customer trend, showing an initial decline, followed by a stable phase, and then another decline in the following years.

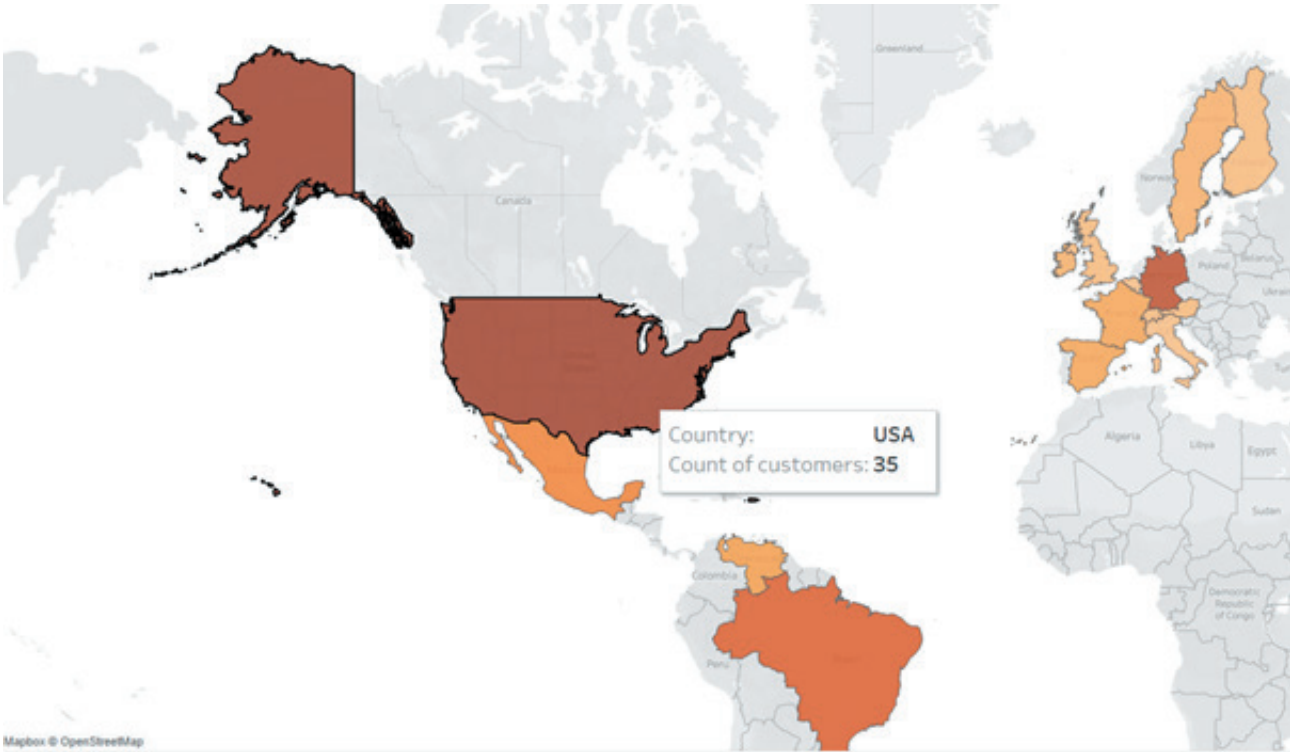


Figure 1. Graph – Customers in Country

Cust..	Customer..	Address	SUM(Price)
2	Ana Truji..	Avda. de ..	
5	Berglund..	Berguvs..	
7	Blondel p..	24 place ..	
11	Bs Bever..	Fauntler..	
13	Centro c..	Sierras d..	
14	Chop-sue..	Hauptstr..	
15	Comérci..	Av. dos L..	
18	Du mond..	67 rue de ..	
20	Ernst Ha..	Kirchgass..	
24	Folk och ..	Åkergata..	
25	Frankenv..	Berliner ..	
30	Godos Co..	C/ Romer..	
33	GROSELL..	5ª Ave. L..	
34	Hanari C..	Rua do P..	
35	HILARIO..	Carrera 2..	
37	Hungry ..	8 Johnst..	
38	Island Tr..	Garden H..	
39	Königlich..	Maubels..	
44	Lehmann..	Magazin..	
46	LILA-Sup..	Carrera 5..	
48	Lonesom..	89 Chiar..	
49	Magazzi..	Via Ludo..	
52	Morgens..	Heerstr. ..	
55	Old Worl..	2743 Ber..	
58	Pericles ..	Calle Dr. ..	
61	Que Delí..	Rua da P..	
63	QUICK-St..	Tauchers..	
65	Rattlesn..	2817 Mil..	
66	Reggiani ..	Strada P..	
67	Ricardo ..	Av. Copa..	
68	Richter S..	Grenzac..	
69	Romero ..	Gran Vía 1	
71	Save-a-Io..	187 Suff..	
75	Split Rail..	P.O. Box ..	

Figure 2. Table – Customer spent

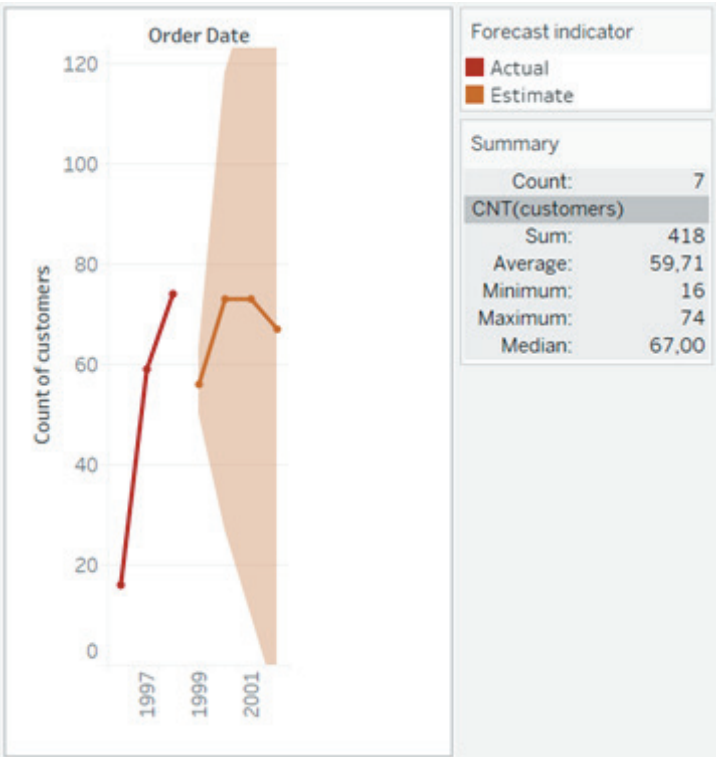


Figure 3. Forecasting



Figure 4. Clustering



On the other hand, Figure 4 demonstrates clustering, which is valuable for customer segmentation, identifying patterns in large datasets, and creating targeted marketing strategies. In this case, clustering is divided into 15 small clusters, where each cluster represents a single country. Essentially, the data is grouped by country, with each country forming a separate unit. When hovering over a circle, a tooltip appears providing clear insights into total profit, number of products sold, and country name. This approach significantly helps users analyse key metrics by specific groups, making it easier to interpret. [6]

3. CONCLUSION

As seen in the provided visualizations, data presented in a visual format is much easier to interpret compared to raw Excel tables. This approach enables faster and more intuitive insights, making it easier to spot trends, patterns, and anomalies. By transforming raw data into interactive dashboards, businesses can make more informed decisions.

Working with SSAS (SQL Server Analysis Services) and SSIS (SQL Server Integration Services) allows for more advanced data management. SSAS helps create multidimensional models, optimizing data analysis and query performance, while SSIS automates data integration (ETL), ensuring data is processed and ready for analysis at scheduled times, without manual intervention. These tools—SQL Server, SSIS, SSAS, and visualization platforms like Tableau and Power BI—work together as an integrated system, improving data processing efficiency and enabling actionable data insights.

For example, data collected from a data warehouse can be processed using distributed systems like Apache Pig or Spark, cleaned, and stored in an accessible format like Excel or an SQL database. SQL is then used for queries, and SSIS and SSAS integrate and model the data for optimized analysis. Finally, Tableau or Power BI visualizes the processed data, making it easier for decision-makers to interpret and act upon. This process illustrates how these tools simplify big data management and enhance decision-making. Effective data visualization and reporting help businesses make data-driven decisions, improve customer relationships, and gain a competitive edge by identifying trends and optimizing resources.

Future research could explore integrating AI and machine learning for predictive analytics, which could lead to more accurate forecasting, personalized customer experiences, and automated decision-making processes, enhancing business intelligence in an evolving digital landscape. [7] [8]

REFERENCES

- [1] “What is business analytics? The complete guide for analysts,” [Online]. Available: <https://www.thoughtspot.com/data-trends/business-analytics>. [Accessed 04 03 2025].
- [2] “What is Business Reporting?,” [Online]. Available: <https://www.jaspersoft.com/articles/what-is-business-reporting>. [Accessed 04 03 2025].
- [3] B. Balusamy, N. R. Abirami, S. Kadry and A. H. Gandomi, *Big Data: Concepts, Technology and Architecture*, Hoboken, NJ: John Wiley and Sons, Inc., 2021.
- [4] D. Liu and V. Y. Yoon, “Developing a goal-driven data integration framework for effective data analytics,” *Decision Support Systems*, vol. 180, p. 114197, 2024.
- [5] A. Ambasht, “Real-Time Data Integration and Analytics: Empowering Data-Driven Decision Making,” *International Journal of Computer Trends and Technology*, vol. 71, no. 7, pp. 8-14, 2023.
- [6] A. Upadhye, “Enhancing business strategy with sales data visualization,” *International Journal of Data Science Research and Development (IJDSRD)*, vol. 2, no. 1, pp. 27-36, 2023.
- [7] M. Mravik, T. Vetrisevi, K. Venkatachalam, M. Sarac and N. Bacanin, “Diabetes Prediction Algorithm Using Recursive Ridge Regression L2,” *Computers, Materials & Continua*, vol. 71, no. 1, pp. 457-471, 2022.
- [8] M. Šarac, M. Mravik, J. Dijana, Š. Ivana, Ž. Miodrag and B. D. Nebojša, “Intelligent diagnosis of coronavirus with computed tomography images using a deep learning model,” *Journal of Electronic Imaging*, vol. 32, no. 2, pp. 1-10, 2022.



STUDENT SESSION

WEB APPLICATION FOR SCHEDULING DOCTOR APPOINTMENTS

Bojan Papaz*,
[0009-0005-0098-3937]

Miloš Mravik,
[0000-0001-5442-3998]

Marko Šarac
[0000-0001-8241-2778]

Singidunum University,
Belgrade, Serbia

Abstract:

This paper represents the design and development of a modern web application that is being used for scheduling medical appointments and aims to solve today's common issues in healthcare systems. The goal of this project is to explore how modern full-stack tools (React, AppWrite, Twilio) can be used to automate appointment booking and communication in healthcare. These issues are missed appointments, scheduling conflicts and inefficient communication. This application combines technologies such as React.js, Next.js, AppWrite and Twilio. This kind of system provides secure login, real-time updates, automated SMS and Email reminders which give information about scheduling, rescheduling and cancelled appointments and role-based access for both doctors and patients. Unlike traditional tools, this solution is fast, scalable and fully automated. The paper also includes a review of similar systems in current research and compares existing approaches in terms of usability, flexibility and automation. The solution that this application gives stands out because the technology and usability make it ideal for clinics and hospitals of all sizes. Future upgrades will include AI-driven scheduling, analytics and telemedicine features. Overall, this application offers a reliable, efficient and research-informed answer for a lot of complicating factors in traditional appointment scheduling.

Keywords:

React.js, Next.js, AppWrite, Twilio, Sentry.

INTRODUCTION

Appointment scheduling is a core process in healthcare, but traditional methods remain slow, prone to errors and are inefficient. Manual bookings often lead to scheduling conflicts, long wait times, and administrative overload, directly impacting both patient satisfaction and provider workflow. Even with the rise of web-based systems many existing tools lack automation, real-time notifications or secure user management. Several recent studies have explored digital scheduling in healthcare. Zhao et al. (2017) conducted a systematic review of web-based medical appointment systems and found that most lacked integration with real-time updates or communication tools. [8] Ele et al. (2020) proposed a solution by sending notifications using SMS, but it lacked user role management and backend scalability. [9] Srinivasan et al. (2024) described a basic doctor appointment app with static scheduling features and limited automation. [10]

Correspondence:

Bojan Papaz

e-mail:

bpapaz@singidunum.ac.rs





These examples demonstrate the need for a more complete, modular, and real-time solution. This paper aims to provide that. This is especially found in clinics with limited technical infrastructure. These limitations create an urgent need for a more modern, efficient and secure scheduling solution tailored to healthcare environments. This paper presents the design and evaluation of a web-based appointment scheduling system built using React.js, Next.js, AppWrite, Twilio, and Sentry. The system was designed to:

- Automate core appointment workflows
- Integrate secure role-based user authentication
- Provide real-time availability updates
- Deliver automated SMS/Email notifications
- Track and resolve system errors proactively
- Support scalable, flexible data structures through a document-based model

The goal is to demonstrate how an integrated, modular design can improve scheduling efficiency, reduce missed appointments and enhance communication in clinical settings. The system is evaluated qualitatively against existing solutions, with a focus on usability, automation and scalability.

2. FRONTEND OF THE MEDICAL APPOINTMENT SCHEDULING APPLICATION

The frontend of the application is developed using React.js and Next.js [2] which provide a responsive and interactive user experience. Frontend uses an architecture utilizing components, allowing modular design and easier scalability. Styling is handled with Tailwind CSS [3] and ShadCN who are enabling the creation of a clean and intuitive UI for patients and healthcare staff. Real-time appointment scheduling is implemented through modal components, such as the one shown in Listing 1, which uses React hooks (useState) and props what are context-aware to dynamically book or cancel an appointment.

```
export const AppointmentModal = ({
  patientId,
  userId,
  appointment,
  type,
}): {
  patientId: string;
  userId: string;
  appointment?: Appointment;
  type: "zakaži" | "Otkaži";
  title: string;
  description: string;
}) => {
  const [open, setOpen] = useState(false);
  return (
    <Dialog open={open} onOpenChange={setOpen}>
      <DialogTrigger asChild>
        <Button
          variant="ghost"
          className={`capitalize ${type === "zakaži" && "text-green-500"} `} >
          {type}
        </Button>
      </DialogTrigger>
      <DialogContent className="shad-dialog sm:max-w-md">
        <DialogHeader className="mb-4 space-y-3">
          <DialogTitle className="capitalize">{type} Termin</DialogTitle>
          <DialogDescription>
            Molimo unesite informacije da {type} termin.
          </DialogDescription>
        </DialogHeader>
        <AppointmentForm
          userId={userId}
          patientId={patientId}
          type={type}
          appointment={appointment}
          setOpen={setOpen}
        />
      </DialogContent>
    </Dialog>
  );
};
```

Listing 1. React Modal Component for Booking Appointments



The React component below demonstrates how TypeScript ensures better readability, type reinforcement and it also reduces potential runtime errors by giving it correct data types.

3. BACKEND

The backend of this application leverages AppWrite, a backend-as-a-service (BaaS) platform to manage core functionalities such as authentication, database operations and notifications. [4] It enables real-time, scalable handling of medical appointment data while maintaining strong security and API integration. AppWrite's SDK is used to initialize the server connection, authenticate securely using environment variables and also interact with multiple services. This includes databases which are document based and interacts with user management. Backend logic is structured to handle operations like creating new appointments, accessing user profiles, and triggering notifications.

As shown above in Listing 2, the backend connects securely to AppWrite services using environment variables for credentials and endpoint configuration. This setup is used to provide a foundation for performing secure CRUD operations on patients, appointments and doctors. Real-time database features and serverless cloud functions are used for automated updates, scheduling conflict checks, and integration with Twilio for notification delivery, which notifies the patient for scheduling, rescheduling and cancelling appointments.

To ensure smooth integration between frontend and backend modules, the system was developed following a modular, full-stack approach. All features were tested in a controlled development environment with attention to security, real-time behavior, and error handling. Core backend features such as authentication, scheduling, and notification workflows were validated through internal testing scenarios simulating real user behavior.

4. APPWRITE

In modern web development, backend infrastructure must support automation, data security, real-time updates and seamless communication with frontend interfaces. This project uses AppWrite as the core backend solution for managing authentication, database operations, file storage and notifications. Unlike traditional backend setups that require manual infrastructure management, AppWrite offers a pre-built yet customizable environment which proves ideal for small to medium-scaled healthcare systems. It also allows fast development cycles, serverless cloud functions, and document-based storage. All these points are key for building scalable and flexible healthcare applications.

4.1. AUTHENTICATION AND USER MANAGEMENT

AppWrite supports multiple secure login methods including email/password and third-party OAuth authentication (for example Google, Facebook...). This ensures flexible and secure access for users with varying technical skill levels whether they are skilled computer specialists or elderly patients. In this system, role-based access control (RBAC) is enforced:

- Patients can register and manage their own appointments
- Doctors can update their schedules
- Admins have access to broader analytics and records

Each user session is protected by JWT (JSON Web) tokens and all data exchange occurs over HTTPS, ensuring that sensitive medical records cannot be intercepted or tampered with. This approach balances usability with security. This system is making it suitable for medical systems that must comply with privacy regulations such as GDPR.

```
import * as sdk from "node-appwrite";

const client = new sdk.Client();
client.setEndpoint(ENDPOINT!).setProject(PROJECT_ID!).setKey(API_KEY!);

export const databases = new sdk.Databases(client);
export const users = new sdk.Users(client);
```

Listing 2. Backend Configuration Using AppWrite SDK



4.2. DATABASE AND APPOINTMENT MANAGEMENT

AppWrite AppWrite uses a document-based NoSQL model that is chosen for its flexibility and ease of scaling, especially in applications like these. This model fits cases where schemas are dynamic and rapid development is essential. This choice was made over relational or graph databases (GDB) because:

- Patient and appointment data can vary significantly across clinics
- The schema needs to evolve without downtime
- Real-time interactions are easier to implement in document-based models

Collections that are used in this applications database include:

- Users - stores patient and doctors profiles
- Appointments - stores active, canceled and completed appointments
- Notifications - tracks SMS/Email logs and statuses

Referential integrity is maintained through unique user IDs, appointment IDs and foreign key references. This prevents issues such as name duplication (for example two patients named “Ana” or more doctors named “Jovan”).

To improve traceability each notification includes metadata such as:

- Date and time sent
- Provider used (Twilio)
- Response status and error codes (if applicable)

4.3. CLOUD FUNCTIONS AND AUTOMATED WORKFLOWS

One of the most powerful features of AppWrite is its support for serverless cloud functions, which allows developers to automate critical backend processes without maintaining separate servers. This is critical in healthcare scheduling, where timing and accuracy are essential. Key use cases include:

- Sending appointment confirmation, rescheduling or cancellation Emails or SMS to patients.
- Triggering reminder notifications 24 hours before an upcoming appointment.
- Freeing time slots when an appointment is cancelled
- Logging and monitoring system activities for security and debugging.

These functions run in an event-driven way and are triggered by changes in the database or user actions. This reduces manual work and improves efficiency across the system.

4.4. SECURITY AND DATA PROTECTION

Since medical applications deal with sensitive patient data, strong security measures are essential to comply with privacy regulations and prevent unauthorized access. AppWrite ensures security in multiple ways:

- Data Encryption – All stored medical records and user credentials are encrypted, ensuring that even if unauthorized access occurs, sensitive information remains protected.
- Role-Based Access Control (RBAC) – Doctors, patients and administrators have different permission levels
- JWT Authentication – JSON Web Tokens (JWT) secure API calls, ensuring that only authenticated users can access certain endpoints.
- Secure API Communication – All backend communication occurs over HTTPS, preventing man-in-the-middle (MITM) attacks and data interception.
- Automated Backups – AppWrite offers regular backups of stored data to prevent data loss in case of failures.

By integrating security best practices, the system provides reliable data protection while maintaining efficiency and accessibility for authorized users.

During development, the system was evaluated based on three primary criteria: usability, automation, and scalability. Usability was assessed through interface clarity and ease of booking appointments. Automation was measured by observing the success of real-time updates and SMS/Email triggers. Scalability was analyzed through the flexibility of the document-based data model and use of cloud functions for asynchronous workflows. These findings were compared qualitatively with other web-based appointment systems from recent literature. The proposed solution showed advantages in notification automation, backend modularity and support for real-time scheduling. All are features that are often missing or incomplete in similar tools.



Table 1. AppWrite Database Structure

Table/Collection	Description	Example Data
Users	Information about users (patients, doctors)	{ "id": "user_001", "name": "Marko", "role": "doctor", "phone": "+381641234567" }
Appointments	Tracks all scheduled, cancelled and completed appointments with linked user IDs	{ "id": "appt_034", "patientId": "user_002", "doctorId": "user_001", "date": "2024-03-20", "status": "Scheduled" }
Notifications	Logs all messages sent using SMS or Email with traceability info	{ "id": "notif_012", "type": "SMS", "status": "Sent", "to": "user_002", "provider": "Twilio", "timestamp": "2024-03-19T12:00:00Z", "responseCode": "200" }
Doctors	Stores doctor-specific info like specialization and availability	{ "id": "doc_005", "userId": "user_001", "specialization": "Dermatology", "availability": ["2024-03-20T09:00"] }

5. TWILIO

Effective communication is a key element in reducing missed appointments and improving efficiency in healthcare. For this reason, the system integrates Twilio, a cloud-based messaging platform, in order to send real-time SMS and email notifications. [5] Unlike manual confirmation systems Twilio enables automated, programmable communication, which improves user experience, reduces administrative workload and also ensures consistent follow-ups with patients. This automation is especially useful in environments where appointment number is high and human error must be minimized.

5.1. APPOINTMENT CONFIRMATION AND REMINDER SYSTEM

The notification workflow is coupled with the appointment scheduling logic. When a new appointment is created, a confirmation message is immediately sent to the patient. Additionally, automated reminders are scheduled 24 hours and 1 hour before the appointment, reducing the likelihood of missed visits. If an appointment is canceled or modified, both the doctor and the patient receive instant updates on the situation. This two-way communication ensures transparency and helps medical staff adapt schedules without any delay. All messages are structured and tagged with metadata including timestamps, delivery status and user IDs for full traceability.

5.2. TWILIO API INTEGRATION AND WORKFLOW

The backend communicates with Twilio's REST API to send notifications efficiently. The workflow consists of:

- A patient books an appointment using the frontend.
- The backend registers the appointment in the AppWrite database.
- A Twilio API request is triggered to send an SMS or email confirmation.
- Automated reminders are scheduled for upcoming appointments.

Using Twilio for automated communication in medical appointment scheduling provides several advantages:

- Reduces missed appointments - Patients receive timely reminders, improving clinic efficiency.
- Enhances patient experience - Convenient communication improves trust and reliability.
- Saves administrative time - Eliminates manual phone calls for confirmation and reminders.
- Scalability and reliability - Can support large healthcare institutions without performance issues.

In addition to SMS notifications, email alerts are implemented to provide additional confirmation, ensuring patients are informed through multiple channels.



5.3. SECURITY AND COMPLIANCE CONSIDERATIONS

All data related to messages is handled securely. Patient identifiers are encrypted, and communication with Twilio's API occurs via HTTPS. The system does not store raw message content, but logs only metadata necessary for analytics and traceability, such as:

- Provider used
- Delivery status ("sent", "failed", "delivered")
- Timestamp of delivery

This design balances privacy with operational monitoring and allows healthcare providers to track communication effectiveness and also troubleshoot delivery issues when they are needed.

A qualitative comparison was conducted between the proposed system and other web-based scheduling solutions discussed in recent literature, including those by Zhao et al. (2017) and Ele et al. (2020). Unlike these systems, which either lack real-time scheduling or manage to provide only basic SMS notifications, the proposed platform offers full automation through Twilio integration, delivering both SMS and email alerts. It also includes role-based access for patients, doctors and administrators. However previous systems implement only basic or static user roles. Real-time scheduling is supported through the combination of React's dynamic frontend and AppWrite's real-time backend database, which is absent in the earlier models. The proposed system integrates Sentry for full error tracking and performance monitoring, which are features not present in prior solutions. Architecturally, it adopts a modular, serverless design using AppWrite cloud functions, offering greater flexibility and scalability compared to the monolithic or semi-modular structures described in earlier research. Finally, its document-based NoSQL database provides schema flexibility and referential integrity, which addresses limitations observed in rigid relational models used by other tools.

Compared to related systems reviewed in Zhao et al. (2017), Ele et al. (2020), and Srinivasan et al. (2024), the proposed platform delivers deeper automation, real-time availability handling and multi-channel communication. It also supports role-based access, full backend modularity and performance monitoring which are elements missing or limited in prior work.

6. SENTRY

In Modern web applications in healthcare require more than just functionality. They demand reliability, stability and real-time monitoring. For this reason, Sentry is integrated into both the frontend and backend of the system to support error tracking, performance monitoring and root cause analysis. [6] Unlike traditional debugging methods, Sentry allows developers to identify issues and resolve them before they affect users. This is particularly important in healthcare platforms where system errors can directly impact appointment scheduling and communication with patients.

6.1. ERROR TRACKING AND DEBUGGING WITH SENTRY

Sentry provides a detailed, real-time log of application errors and their context. Whenever a bug or exception occurs, like a failed database write or frontend component crash, Sentry captures:

- The type of error and exception message
- The environment in which the issue occurred (production or development)
- Device, browser and Operating System details
- The sequence of user actions leading to the error

This data is then visualized in the Sentry dashboard which allows developers to quickly identify, reproduce and resolve issues. It reduces downtime and prevents critical failures in scheduling or notifications.

In the backend, Sentry monitors server-side events like API failures, Missed notifications and Database connection errors.

6.2. PERFORMANCE MONITORING WITH SENTRY

Beyond tracking crashes, Sentry also provides performance monitoring. This feature allows developers to:

- Analyze application performance on different devices and platforms
- Identify slow and inefficient parts of the application
- Measure API response times and detect bottlenecks

This allows for optimizations driven by data, which are improving the experience for both patients and doctors. For example, if appointment booking pages begin to load slowly, developers can trace the issue to a specific component or API call and address it proactively.



6.3. SCIENTIFIC AND PRACTICAL VALUE OF SENTRY INTEGRATION

The integration of Sentry adds scientific and operational value by supporting quantitative analysis of system stability, providing reproducible debugging data for case studies and enabling ongoing quality assurance with minimal developer intervention. In scientific research terms, this aligns with system reliability engineering and software quality assurance principles. It ensures that the solution remains usable and scalable as more users interact with the platform.

7. CONCLUSION

The primary goal of developing this web application was to improve user experience, enhance efficiency in healthcare institutions, and reduce errors associated with traditional appointment scheduling methods. Digitalization of this process is becoming increasingly essential not only for large clinics but also for smaller medical offices that want to optimize their operations. The system successfully provides a user-friendly and efficient platform for scheduling and managing medical appointments, ensuring accessibility for both patients and healthcare providers. The React.js and Next.js frontend deliver a smooth and intuitive user experience, while the AppWrite backend ensures secure authentication, data management and real-time updates. The integration of Twilio for automated notifications and Sentry for performance monitoring further enhances system reliability. Future enhancements could include pagination for handling large amounts of data efficiently, improving user navigation with breadcrumbs and refining form design for better accessibility. Additionally, implementing Two-Factor Authentication (2FA) would significantly improve security by adding an extra layer of user verification. [7] Overall, this web application represents a modern, scalable, and secure solution for appointment scheduling in healthcare institutions. Its implementation reduces administrative workload, minimizes scheduling errors, and improves communication between patients and doctors. With future improvements and optimizations, it has the potential to become an even more robust and indispensable tool for medical institutions seeking to modernize their services. The system was designed and evaluated to solve key challenges in appointment scheduling. Compared to existing tools, it demonstrates strong advantages in automation, real-time functionality, and data architecture.

Future enhancements such as AI-powered scheduling and improved analytics will further increase its value across clinics. The system was tested across 3 roles (Admin, Doctor, Patient) in a controlled localhost setup. Booking, cancelling and notifications were tested with real data. Checks confirmed real-time updates and notification delivery. As for the core flow, no issues were observed.

REFERENCES

- [1] R. Team, "React.js Documentation," Meta, 25 January 2024. [Online]. Available: <https://react.dev/blog>. [Accessed September 2024].
- [2] T. Labs, "Tailwind: A Utility-First CSS Framework," Tailwind Labs, 18 March 2024. [Online]. Available: <https://tailwindcss.com/docs/installation>. [Accessed September 2024].
- [3] A. Team, "AppWrite Documentation," AppWrite, 5 April 2024. [Online]. Available: <https://appwrite.io/docs>. [Accessed September 2024].
- [4] T. Inc., "Twilio API Documentation," Twilio Inc., 10 February 2024. [Online]. Available: <https://www.twilio.com/docs>. [Accessed September 2024].
- [5] S. Docs, "Sentry: Error Tracking and Performance Monitoring," Sentry, 30 January 2024. [Online]. Available: <https://docs.sentry.io/>. [Accessed September 2024].
- [6] R. Aghili, Q. Qin, H. Li, F. Khomh, Understanding Web Application Workloads and Their Applications: Systematic Literature Review and Characterization," *arXiv preprint*, 2024.
- [7] M. Šarac, M. Mravik, D. Jovanović, I. Strumberger, M. Živković, N. Bačanin, "Intelligent diagnosis of coronavirus with computed tomography images using a deep learning model," *J. Electron. Imag.*, vol. 32, no. 2, p. 10, 29 November 2022.
- [8] P. Zhao, I. Yoo, J. Lavoie, B. J. Lavoie and E. Simoes, "Web-Based Medical Appointment Systems: A Systematic Review," *Journal of Medical Internet Research*, vol. 19, no. 4, p. 134, April 2017.
- [9] B. I. Ele, J. A. Odey, N. E. Frank, I. M. Ekinya, "A Web-Based Medical Appointment Scheduling with SMS Alert Notification System," *Transactions on Machine Learning and Artificial Intelligence*, vol. 8, no. 6, pp. 28-38, December 2020.
- [10] S. Srinivasan, N. Duraimutharasan, D. Ganesh, S. Nagaraj, R. S. Lakshmi Balaji, "A Web-Based Doctor Appointment System," *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 11, no. 2, p. 8, February 2024.



THE IMPACT OF ARTIFICIAL INTELLIGENCE IN MARKETING

Anđela Pavlović*,
[0009-0004-0548-8778]

Marko Šarac
[0000-0001-8241-2778]

Singidunum University,
Belgrade, Serbia

Abstract:

The core purpose of this paper is to examine elements that play a role in influencing Artificial Intelligence tools and algorithms with their implementation. The paper examines the impact of Artificial Intelligence on marketing, trends, and factors that have an effect on implementation and innovations. Topics include digital marketing, big data, ethical concerns and transparency in AI applications.

The paper is divided in two parts. In the first section, we are going to explain essential concepts and fundamentals correlating AI algorithms and models on marketing. An analysis is looking into which factors play a role in decision-making using AI in marketing. The second part investigates uses and trends in AI-driven marketing. Examples from practice are provided to show how artificial intelligence is applied in marketing in various industries and different ways.

Keywords:

Artificial Intelligence, Marketing, Ethics, Algorithm, Data.

INTRODUCTION

Since marketing is something that surrounds us, it's no surprise that the hand of artificial intelligence grabbed upon it. Marketing is a versatile industry and a sphere of study that has a base in psychology and creativity. Marketing has a duty in providing customers with satisfaction, as in with various activities to promote different products. Artificial intelligence is widespread. The first thing that associates us with it is usage in a creative way. AI already has effective and positive results linked with marketing. AI is based on machine learning and deep learning with the help of human input. AI is still not in daily use, but it is not discriminated against. A lot of risks are still to understand and people are being very cautious. Often the reason it is avoided is due to not having a proper understanding of it and fear of the unknown. In some practices, it is not used because people are scared that it would be too obvious that it was used. Another reason is that AI has not explored its full potential yet.

Correspondence:

Anđela Pavlović

e-mail:

andjela.pavlovic.21@singimail.rs



Other risks include privacy concerns and ethical regulations. For a company to have a data breach, especially one that manages sensitive user information, would be catastrophic, not only for the users, but for the company and their reputation. Even the data that is fed to the AI must be carefully filtered.

Marketing has many roles within a company, such as attracting customers, and achieving organizational goals with a focus to be the customers. The aim is to reach a widespread audience. Even with all the positive results, the risks are sometimes too high. Smaller companies can't take chances with the investment risk of implementing AI in their marketing solutions. They need to do further analysis and look at the wider picture before making any decisions. For example, Alibaba used AI to create text for advertisements. It is fully autonomous, without need for human input. AI chatbots are another example of AI being used in a safe way that has already been evaluated. They are taught to respond to customers' questions. AI is needed to be used when it comes to decisions that are data based, especially when it comes to understanding customers' needs and wants. The results have a positive effect on the sales. But blatantly using AI will not solve all the problems, especially when customers like personalized offers. AI is also used when it comes to setting up the prices for various products, while doing analysis of the market. The positive side of AI is that with nonstop use of it, you figure it out and gain expertise in it. [1]

Another instance of employing AI can be illustrated on the example of Coca Cola. Coca-Cola incorporates AI in marketing for personalized advertising campaigns, social media sentiment analysis and predictive analytics for pricing and consumer behavior. These applications have enhanced consumer engagement and optimized marketing strategies. [2]

The point relates to maximizing sales, profit and customer engagement, using artificial intelligence to ease the work is a normal and expected response. The writer pointed out the fact that AI can hold prejudice against some ethical groups. Some data and information are not accessible which leads to informational asymmetries. Customers often want to see emotions when talking to AI or else they will be unhappy. Newer models of AI have the ability of personalization for customers which happens to be one of the upper hands of AI marketing. AI also has the possibility to offer personalized offers which can be adapted to individual needs. Companies depend on artificial intelligence to analyze datasets that would take too much time for a person to do it. Psycho-

social traits are one of the patterns that can be predicted from those sets and used to create content catered to us. [3]

Overall, AI has been especially useful in marketing and industries that can afford it. By using it carefully and in the ways that have been previously proved as good practice, results are guaranteed to be positive. However, the effectiveness of AI-driven strategies depends on ethical implementation and regular updates. While AI offers powerful insights and automation, businesses must manage the risky side of it, consumers trust, ensuring transparency and data privacy. As technology evolves, companies that use AI have the possibility to gain competitive advantage while fostering long-term customer relationships.

2. FACTORS THAT PLAY A ROLE IN AI MARKETING

When implementing AI in marketing, a few components are in the focus, data collection being one of them. Large volumes of data are gathered from various sources, for example, customers interaction, social media engagement and purchase records. This data is called big data and is the core of AI marketing. Customer profiling is when AI analyses collected data for a better understanding of a customer profile. Then, a customer could be segmented into a specific group catered to their demographic characteristics. Machine learning algorithms are used to identify patterns and predict the future based on big data. Behavioural targeting is used when AI examines real-time user interactions and delivers personalized content and advertisements at the right moment to maximize engagement. By learning user behaviour, AI refines its targeting, making marketing strategies more effective and improving customers' experience. Predictive analytics continuously runs throughout the entire marketing process. A/B testing is a method known as split testing used to compare, for example two types of marketing ads to determine which one performs better. It is a useful optimization tool for making informative decisions. Cross-channel integration utilizes personalized AI content on different channels, for instance, social media platforms and web sites. Ethical considerations are a risky part of AI marketing strategies. Privacy regulations and user consent must be the focus when building trust. Key elements crucial for successful AI personalization are data quality and quantity, user privacy, user engagement, cultural sensitivity and others. [4]



Rapid advances in AI marketing could risk brand messages to be identical with the lack of human creativity. While AI can analyse vast amounts of data, it can not understand the emotional factors that play the role in customer behaviour. For example, to counter biases that AI has been trained on or spotted patterns in datasets, few organizations within the same industry would collaborate and share their representative datasets. That way, AI enhances fairness and efficiency. [5]

On the other hand, using generative AI can have positive effects on the marketing team. It can enhance creativity and productivity leading to better outcomes and improved marketing processes. Personalization is essential for creating more engaging content and capturing a wider audience's interest. [6]

When talking about understanding the algorithm of AI, we face transparency issues. AI algorithms operate in a way that is not easily understood. Based on inputs, AI produces outputs, but the reasoning and logic behind decision-making processes is unclear and unknown. Opaque algorithms are a type of algorithm in which users see the results but have no insight as to why the AI made these choices. This lack of transparency can lead to multiple issues with trust concerns, biases, regulatory and ethical problems. In marketing, having a bias or fairness complications can cause severe complications with the output and customer satisfaction. To ensure users are aware of what data is collected and how it's used, businesses opt to implement transparent privacy policies communicating about AI algorithms. Customers are allowed to choose which data can be collected. Providing them with this choice and control enhances their comfort. Nowadays companies are investing in algorithm accountability, monitoring if biases happen and solving challenges that emerge. This contributes to organization making sure that their AI systems are working properly assuring customers loyalty and trust. Marketers focus on explaining how AI operates and how it is incorporated into their strategy-driven decisions. AI-driven tools can help with automatization of repetitive tasks, making employees in marketing industry focus on higher-value tasks and strategies that can elevate their brand message. [7]

As shown in Table 1, there is a clear difference between traditional marketing and AI marketing. While traditional marketing relied on strategies and methods that are now considered inefficient and in outdated, AI marketing represents the future. Content personalisation refers to marketing messages and content based on personal preferences of a customer. In traditional marketing it used to be limited to broad segmentation. That means it was based on demographic data or geographic location. Broad segmentation means that marketing strategy was catered to a larger group of customers. If we import AI in this feature, results are highly customisable using AI algorithms, which indicates they are based on individual needs, preferences and wants. Data analysis used to be slow in traditional marketing, because it was based on manual methods. Marketers would collect data over time and then look through it. Since it was based on a human factor, it was time consuming and prone to errors. AI uses real-time automated processing and delivers results in no time. It can spot patterns immediately and organise results in an comprehensive manner. Cost efficiency in traditional marketing was expensive due to a few factors such as manual labor, different expensive channels and limited targeting. With the rise of automation, cost efficiency improved. Now AI can automate the tasks that used to be hand-operated, improve targeting and enhance marketing campaigns. Return on investment in traditional marketing was difficult due to issue related to tracking effectiveness of marketing campaigns. In AI marketing, ROI measurements are clear and precise, because of the algorithm that collects data and instantly provides insights.

Considering all factors, the implementation of AI in marketing can bring numerous benefits from customer profiling to boosting effectiveness of brand messages. With the help of machine learning and predictive analytics, businesses can improve their strategies and customer engagement. Regardless of positive aspects, that can't go without addressing problems like transparency, ethics and discriminations against social groups so the trust and loyalty can be built.

Table 1. Traditional Marketing vs. AI Marketing

Feature	Traditional Marketing	AI Marketing
Content Personalisation	Limited, broad segmentation	Highly customized using AI algorithms
Speed of Data Analysis	Slow, manual analysis	Real-time automated processing
Cost Efficiency	High campaign costs	Optimized through automation
ROI Measurement	Difficult	Precise through AI analytics



On the one hand, AI can boost productivity, cut time in repetitive tasks using powerful tools, but on the other hand human touch and emotional connection can't be that easily replaced with full automation. By focusing on ethical practices and perfecting AI tools, along with other areas where it can provide assistance, AI can reach its fullest potential. With the development of AI, marketers must balance technological advancements with the principles of creativity, fairness, and clarity to create valuable and influential customer experiences.

3. EMERGING TRENDS AND APPLICATION IN MARKETING

With technology advancing, it's only natural for it to have application within various industries, including marketing. With continuous advancements of AI, new trends are rising. AI adapts to be used and personalized for different needs. Different ways of implementing AI is working in favor of businesses, cutting costs and time and increasing market rivalry.

AI has been used in green marketing. Green marketing is marketing that promotes practices in business that are environmentally friendly. The point is to attract customers who prefer sustainable options while caring for the environment and reducing the negative practices. "Green AI" focuses on environmentally friendly algorithms with the goal to minimize negative impact of AI on environment. This message fits customers wishes for brands that promote sustainability and only in products, but in technology too. AI also can optimize inventory, reduce waste and have a role in supply chain operations. [8]

Natural language processing is a part of AI that plays a big role in digital marketing. It is used to interpret human language. It can be applied in sentiment analysis, which is meant to evaluate clients' emotions. NLP is used in chatbots. Chatbots became a practice that is widely adopted, because of their capabilities to respond to inquiries, offer personalized suggestions and responding to questions. In return, chatbots gain data about customer preferences, which can be enforced in marketing strategies. Computer vision is a branch of AI that enables the analysis of images and videos to identify various elements. It can recognize different products in visual content and deliver targeted advertisements based on these items. Another way to use AI is in its generative form. It can produce text, photos and videos. This tool enhances productivity and helps marketers focus on strategy and reduces time and resources. AI can also be used in dynamic pricing. AI algorithms adjust prices

according to elements that influence them, such as demand, competition and other economic factors. Using algorithms for pricing allows companies to reach their audience easier and improve their return on investments. [9]

Following rapid advances, AI can be implemented in process of web design. It's possible to create sites that are engaging and user-friendly. This integration helps with the automation of web design tasks, including layout designs paired with content creation, which cuts the costs of website development. AI algorithms improve user satisfaction analysing website performance and user habits. [10]

Neuromarketing explains how the brain reacts to advertisements and brand-related content, combining neuroscience and psychology. The information gained this way can be further explored by AI. AI can identify which products are customers more likely to purchase and parts of brain influencing this decision can be revealed. [11]

Artificial intelligence plays a role in forecasting trends. Using this knowledge, marketers can optimise their budget and further analyse corresponding strategy. Any company that implemented AI in their marketing system can gain competitive advantage and their conversion of customers to be loyal has a higher chance. For the positioning of the brand and its message, AI that gathers data is appreciated. Thanks to different data, starting from demographic, such as age, gender, location to other data, like persons interests, the best target market for the brand is found. Marketers are using programmatic advertising, which is based on machine learning and real-time events. In that way, advertisements are catered to their ideal customers. AI helps minimise mistakes, as long as marketing specialists who are working with it provide monitoring and control. [12]

Using AI in marketing is one of the best practices available, but not everyone would agree. Businesses now face a problem into convincing stakeholders to invest in AI systems. To prove a point, tracking ROI is obvious, but showing how AI bettered user-satisfaction is hard. Stakeholders are scared of facing legal issues and damage that could happen to their image. [13]

There are a few hybrid AI models used for omnichannel personalisation. Omnichannel personalisation refers to the use of different communication channels, online and offline to provide personalised customer experience. Data collected from various channels is used to adjust content to customers taste. Multi-Modal AI combines text, image and engagement data to deliver



products that cater to customers preferences. Federated Learning is used for AI training. The data that is used in this training is decentralised, which means that if it comes to privacy breaches, most likely it won't affect personal information, since it is stored on different systems. Edge AI is technology that processes data locally and not on a cloud. That results in responses to be fast and effective. Data that is used in marketing comes from five sources. First-Party data is data that comes directly from customers interactions. Second-Party data is data that is shared between companies that work together. Third-Party data is external and comes from trends and social-media interactions. Sensor and IoT data is data that comes from different smart devices. Social media and sentiment analysis refers to AI analysing social media engagement. [14]

Including AI in CRM leads to automatization in sections such as email marketing, segmentation of customers and predictive analysis. AI cuts time and repetitive work in A/B testing and audience targeting. It achieves the possibility to forecast market changes before making a decision on marketing promotions. The main problem with including AI in email marketing is the authenticity of the message that is being broadcast to customers. AI can copy the way humans write, but can't express emotions. Other problems include plagiarism and issues with replacing human labor. The rise of AI is followed by integrations with augmented reality and virtual reality. [15]

Augmented reality is a technology that uses smart devices in a way that digital elements overlay onto the real world. It combines real-world environments with generated images and sounds. AR is nowadays combined with AI in a way that enhances personalization and interactive experiences. In tourism marketing, AR and AI can bring together a captivating experience. AR is increasingly being used in integrated marketing communications. With the help of augmented reality, all messages look more put together and bring a unique customer experience that makes a brand stand out. By integrating AI, customers feel more satisfied. [16]

With all this said, AI is emerging to be more and more used in marketing, in all processes. With development of technology, it is expected of AI to be even more important in the future, enabling faster adaptation to market changes. By implementing AI tools, businesses can enhance customers experience and improve competitiveness.

4. CONCLUSION

The core purpose of this paper is to research elements influencing AI tools and algorithms with their implementation in practice. The vast use of AI is helping marketers and business be cost-effective and efficient. To rely only on artificial intelligence with content creation and handling customers inquiries, can be a bad practice, since people want to see that human touch. On one hand, productivity and creativity of marketing team could be boosted, but on the other, human interaction and imagination is needed for the brand message to be unique, differentiating it from their competitors.

Different elements have a role influencing AI algorithms. One of them is big data which is a core factor. Without data, patterns, user-behaviour, customer profile and personalization couldn't be made. When collecting data, privacy policy and user agreement play a big role. Being transparent with data that is analysed can build trust with customers. AI in marketing presents ethical concerns, including data privacy and transparency. Many AI algorithms operate as "black boxes," where the logic behind decisions remains unclear. Ethical AI implementation requires:

- Transparent data collection policies
- Continuous algorithm monitoring to prevent biases
- Consumer control over shared data

For AI to be ethical in marketing, it's needed for consumers to be informed about data collection and provide their consent before data is acquired. Then, data collected should be limited to what is necessary and de-identified to make sure that the privacy of consumers stays protected. That data needs to be behind a lot of strong security protocols so the data-breaches are unable to gain access to it. Another issue that has to be resolved is AI black-box algorithms that generate output without providing clear context or revealing what had an impact on their decision. Algorithms providing outcomes should have straightforward explanations of their functionality, resulting in customers gaining trust in handing over their data. If consumers understood the logic behind targeted ads and personalized offers, that would influence their decision-making to be more knowledgeable. [17]

It is needed to constantly monitor and polish AI algorithms to ensure they remain accurate and productive. Following up with the feedback, adjusting data can help AI models to stay up to date. While AI can have



a positive impact, it is needed to balance its use with human engagement to maintain authenticity and avoid repetitions that could be caused relying on tech. By combining the power of AI with human insight, business can reach new highs of marketing efforts and build stronger, more lasting relationships with the customers. AI is revolutionizing marketing, improving efficiency and personalization. However, a balanced approach is essential, integrating AI tools while maintaining human creativity and emotional connection. Future advancements should prioritize ethical AI usage, ensuring trust and fairness in AI-driven marketing strategies.

REFERENCES

- [1] S. Turkyilmaz, "AI in Marketing," in *Generative AI for Transformational Management*, IGI Global, 2024, pp. 211-240.
- [2] M. Šarac, *Internet Marketing*, Singidunum University, Belgrade, 2015.
- [3] S. Wang, "The Influence of AI in Marketing," *Advances in Economics, Management and Political Sciences*, vol. 52, pp. 52-57, 2025.
- [4] D. R. Deepa, "A Study on the Effectiveness & Factors Influencing Personalized AI-Driven Marketing Among Various Customers – Concerning Chennai City," *2nd National Conference on Future Landscape of Sustainable Business Practices in AI World (NCS-BP'24)*, Rajalakshmi Engineering College, 2024.
- [5] R. Ejjami, "Holistic AI-Enhanced Marketing Framework Theory: Bridging Human Creativity and AI for Ethical Marketing," *International Journal For Multidisciplinary Research*, 2024.
- [6] C. S. Du Plessis, "Beliefs and Adoption of AI in Content Marketing: Insights from South African Marketing Agencies," *South African Journal of Information Management*, 2024.
- [7] A. K. Sharma, "Emergence of AI in Marketing and Its Implications," *International Journal of Marketing and Technology*, 2024.
- [8] D. J. B. Agustawan, "Leveraging AI to Enhance Green Marketing Strategies," *Jurnal Ekonomi Manajemen Akuntansi dan Keuangan*, 2025.
- [9] M. Pattanayak, "Optimizing Digital Marketing with AI," in *Mind and Machines: The Psychology of Artificial Intelligence*, Jyotikiran Publication, Pune, 2024, pp. 165-177.
- [10] R. Ahuja, "Evaluating the Integration of AI in Digital Marketing Automation," *African Journal of Biomedical Research*, 2024.
- [11] S. Agarwal, "The Future and Impact of AI in the World of Marketing," *International Journal of Research and Analytical Reviews*, 2024.
- [12] I. Towhidul, "Artificial Intelligence (AI) Applications for Marketing: Real-Time Analytics in Banking & Finance," in *Proc. Int. Conf. Artif. Intell. Educ. (ICAIE-25)*, Tokyo, Japan, 2025.
- [13] K. D. Kumar and M. Gurusamy, "The Application of Artificial Intelligence (AI) in Marketing: Opportunities and Challenges," *Indian Journal of Natural Sciences*, vol. 14, pp. 54258-54263, 2023.
- [14] A. Dave, "Integrating AI-Driven Personalization in Enterprise Marketing Applications," *Proceedings of the International Conference on AI in Marketing*, 2024.
- [15] F. V. Emmanuel, "The Rise of AI-Generate Content and Marketing Automation," unpublished manuscript, Obafemi Awolowo University, Mar. 2025.
- [16] B. Setiawan, "The Augmented Reality (AR) Based on Artificial Intelligence (AI) in *Integrated Marketing Communication (IMC) in Tourism Villages*," Media Wisata, 2024.
- [17] B. B. Siham, "Ethical Algorithms: Balancing the Power of AI in Marketing with Consumer Trust - Case Studies of International Experiences," *Journal of Business Administration and Economic Studies (JB AES)*, 2024.



A COMPARATIVE STUDY OF OBJECT DETECTION ALGORITHMS FOR SECURITY APPLICATIONS

Roman Kriuchkov,
[0009-0005-9880-3614]

Timea Bezdan*
[0000-0001-6938-6974]

Singidunum University,
Belgrade, Srebia

Abstract:

Currently, there are many different computer vision models available, and each has its unique characteristics. Selecting the most suitable and, importantly, well-performing model can be challenging for companies and researchers who plan to use artificial intelligence to solve their problems. This study aims to evaluate the performance of four prominent computer vision models: YOLOv5, Faster R-CNN, SSD 300, and RetinaNet. The models were assessed on their ability to detect and classify weapons in images. The primary metrics used for evaluating their performance are mAP@50 and mAP@50-95. The dataset used for testing these models is taken from the well-known dataset platform Kaggle and consists of images of various types of weapons sorted by class. This circumstance also makes it possible to associate this research with the field of security and its automation. Experimental results identified YOLOv5 as the best-performing model among the four. The overall performance was constrained by the dataset's limited size and image quality, with the highest mAP@50 reaching 0.8. The findings of this study offer practical insights for companies seeking effective computer vision solutions, as well as for researchers examining the development and comparative performance of object detection models.

Keywords:

Artificial Intelligence, Computer Vision, Classification, Object Detection, Security Applications.

INTRODUCTION

The field of artificial intelligence, specifically computer vision, has become one of the most active and impactful areas of research in recent years. Computer vision technologies enable the automation and simplification of numerous tasks across various domains[1]. This study focuses on a key capability of this technology, object detection in images, a fundamental task in computer vision. In addition to detecting objects, the models are also evaluated on their ability to classify them into relevant dataset categories. Various types of weapons, including both firearms and edged weapons, served as the objects to be identified.

Correspondence:

Timea Bezdan

e-mail:

tbezdan@singidunum.ac.rs



The ability to promptly recognise dangerous objects is both critical and highly valuable[2]. There are numerous environments where enhanced security is essential, including airports, schools, hospitals, and ports[3]. In most of these locations, security is managed by humans, which inherently introduces risks associated with human factors such as fatigue, distraction, or even negligence[4]. Computer vision-based tools in this area mainly aim to enhance security, reduce the likelihood of human errors, and improve monitoring capabilities for video and image data obtained from surveillance cameras or other information transmitters[5].

Successfully performing all the tasks outlined above requires artificial intelligence systems to be trained on a properly constructed dataset and based on a high-quality model capable of effective training and demonstrating high performance in accomplishing the given tasks. Consequently, conducting a comparative analysis of state-of-the-art models is essential to determine the most suitable candidates for such applications. Despite the availability of numerous object detection models, their comparative performance in real-world security contexts, especially with limited-quality datasets, remains underexplored. This raises the key research question: Which object detection model performs best in identifying and classifying weapons in images under constrained data conditions? This research compares four popular and widely used computer vision models, YOLOv5 [6], R-CNN [7], SSD 300 [8], and RetinaNet [9], to identify the most efficient one, defined as the model demonstrating the highest accuracy in object detection tasks. The models' performance is evaluated based on two key metrics mAP@50 [10] and mAP@50-95 [11]. Additionally, the study outlines the dataset preparation process for model training, emphasizing its critical role in the successful application of artificial intelligence in computer vision.

This work contributes by providing a practical comparison of these models on a real-world dataset involving weapon detection, offering insights into their suitability for security applications under constrained data conditions.

The rest of the paper is organised as follows: Section 2 describes the methodology. Section 3 presents the experimental setup, the dataset used, and the results along with their analysis. Section 4 concludes the paper by summarising the key findings.

2. METHODOLOGY

The computer vision models used in this research have different architectures. For instance, YOLOv5, SSD 300, and RetinaNet are single-stage detectors, whereas Faster R-CNN is a two-stage detector. The primary difference between these two approaches lies in the number of processing stages required for detecting and classifying objects within an image: a single-stage detector needs just one pass, while a two-stage detector requires two. During the first stage, region proposal networks (RPN) are generated, and region of interest (ROI) pooling is performed to extract features for each candidate region. The second stage in this architecture is responsible for classifying detected candidates and refining their bounding boxes. As a result, the two-stage detector requires more processing time per image compared to single-stage models, which can be a critical consideration for real-time or resource-constrained applications.

Single-stage object detectors generally follow a modular architecture comprising three main components: (i) the backbone, (ii) the neck, and (iii) the head. While they share this overall structure, the specific implementations can vary significantly across models at both high and low abstraction levels. For models such as YOLOv5, SSD 300, and RetinaNet, the backbone is the main convolutional neural network responsible for feature extraction from the original image. As a result, the backbone identifies both high-level and low-level features. High-level features include shapes and objects, while low-level features correspond to edges and textures [12].

To combine low-level and high-level features, a component known as the neck is used. The neck is an additional set of layers, which acts as a bridge between the backbone and the head, aggregating and refining feature maps from different levels to enhance detection performance. This component creates multi-scale image representations and is implemented in YOLOv5 and RetinaNet. However, YOLOv5 uses two algorithms for its neck, Spatial Pyramid Pooling – Fast (SPPF) [13] and Path Aggregation Network (PANet) [14], whereas RetinaNet employs the Feature Pyramid Network (FPN) [15]. The SSD 300 model does not have a neck component as such, instead, it adds extra convolutional layers after the backbone. These additional layers create multi-scale feature maps, enabling the detection of objects of varying sizes.

The final component in these models is the head, which generates predictions based on the feature maps received from the neck. It is responsible for producing the model's outputs, including object classifications and



bounding box coordinates. Although the implementation of the head varies across models, they all incorporate a multi-scale prediction mechanism, which allows the model to make predictions at different feature levels. This approach enhances the detection of objects of varying sizes by leveraging both fine and coarse spatial information.

Another distinguishing feature among the four models is the resolution of the input images they process. YOLOv5 uses a default resolution of 640×640 , while SSD 300 operates on 300×300 pixel inputs. RetinaNet and Faster R-CNN use variable input sizes, typically with the shorter side resized to 800 pixels and the longer side not exceeding 1333 pixels. Further details regarding each model's configuration are provided in Section 3.1, Experimental Setup.

As mentioned earlier, the metrics mAP@50 (Mean Average Precision at 50% IoU) and mAP@50–95 were used to assess the performance of each model. These metrics evaluate how accurately a model detects and localises objects based on the overlap between predicted and ground-truth bounding boxes. mAP@50 measures the average precision when the Intersection over Union (IoU) threshold is fixed at 0.50. In contrast, mAP@50–95 is a more comprehensive metric that calculates the mean of average precision scores across ten IoU thresholds, ranging from 0.50 to 0.95 in increments of 0.05, thereby offering a more rigorous assessment of model performance.

Additionally, precision and recall metrics are reported for each model in this study. Precision indicates the proportion of correctly detected objects among all detected instances, while recall measures the proportion of correctly detected objects relative to the total number of ground-truth objects in the dataset.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The experiment conducted in this study aimed to evaluate and compare the performance of each model on the available dataset. All experiments were implemented in PyTorch and executed using an NVIDIA T4 GPU.

3.1. DATASET PREPARATION

The dataset used in this research, titled "Weapon Detection Dataset", was sourced from Kaggle and contains 714 images of weapons belonging to nine distinct classes: Automatic Rifle, Bazooka, Handgun, Knife, Grenade Launcher, Shotgun, SMG, Sniper, and Sword. In addition to the images, the dataset includes a description file with metadata and YOLO-format annotations stored as plain text files. The dataset was initially divided into training and validation sets in an 80/20 ratio.

During dataset preparation, extensive work was performed to identify and remove duplicate images and to prevent data leakage from the training set to the validation set. Additionally, many classes in the annotations were modified, as they did not match the classes described in the accompanying file and incorrectly represented the objects shown in the images. The final step in dataset preparation involved creating COCO annotations based on the corrected YOLO annotations. This step was necessary to train the SSD 300, RetinaNet, and Faster R-CNN models since they cannot work directly with YOLO annotations.

Figure 1 shows one example of the images taken from the dataset, displaying bounding boxes and class IDs after completing all the data preprocessing steps



Figure 1. Sample image with bounding boxes from the dataset



described above. As a result, 500 images remained in the dataset, with 88% allocated for training and 12% for validation.

3.2. EXPERIMENTAL SETUP

Each of the four models was trained for 75 epochs. The batch size and number of data loading workers were kept consistent across all models, while the learning rate and input image resolution varied depending on the model architecture and its requirements. Table 1 summarises the key training parameters, including the image sizes that produced the best mAP@50 and mAP@50-95 results for each model.

It is also worth noting the significant difference in configuring SSD 300, Faster R-CNN, and RetinaNet compared to YOLOv5. The latter offers a much simpler and more compact configuration process.

3.3. EXPERIMENTAL RESULTS AND DISCUSSION

This subsection describes the results of experiments conducted using all four computer vision models on a dataset containing images of various types of weapons. Table 2 presents the comparative analysis.

Based on the comparative analysis results, the model that demonstrated the best performance was YOLOv5. The mAP@50 metric was equal to 0.8, and mAP@50-95 reached 0.57, indicating strong object localisation and classification performance. Precision of 75% and recall

of 77% suggest that the model occasionally makes mistakes in identifying the desired objects in images. However, these metrics are very close to each other, indicating that YOLOv5 has indeed learned to reliably detect target objects. Given that recall is at 77%, the quality of predictions remains high, meaning the number of false positives is relatively low. This is further supported by the small difference between precision and recall, only 2%. It is worth noting that the model ultimately manages to detect certain types of weapons in images quite successfully, thus fulfilling its intended purpose. Additionally, the model could be further fine-tuned using additional data to improve its results.

Figure 2 illustrates the confidence points in the YOLOv5 model's predictions, as well as boundary boxes. As can be seen in the image, the model can quite reliably detect classes such as SMG, sword, and handgun, with confidence scores ranging from 0.8 to 0.9. However, YOLOv5 still struggles to accurately identify classes like shotgun and knife, with confidence points ranging from 0.6 to 0.7.

The second-best performing model was Faster R-CNN. This model demonstrated results close to YOLOv5 with metrics mAP@50 and mAP@50-95 equal to 0.75 and 0.53, respectively, which are satisfactory outcomes. However, the model's precision (53%) and recall (60%) were considerably lower than those of YOLOv5. Due to these low precision and recall scores, the model is likely to perform poorly in detecting weapons within images, and it may be less reliable in real-world weapon detection tasks.

Table 1. Parameter configuration

Model	Batch size	Learning rate	Workers	Image size (pixels)
YOLOv5	4	1e-2	2	640 (long side)
SSD 300	4	1e-4	2	300x300
RetinaNet	4	1e-4	2	512x512
Faster R-CNN	4	1e-4	2	800 (short side), max 1333 (long side), aspect ratio preserved

Table 2. Comparative analysis

Model	mAP@50	mAP@50-95	Precision	Recall
YOLOv5	0.80	0.57	75%	77%
SSD 300	0.57	0.36	36%	52%
RetinaNet	0.38	0.20	20%	48%
Faster R-CNN	0.75	0.53	53%	60%



Figure 2. YOLOv5 predictions for validation dataset

Table 3. Per-class evaluation metrics for YOLOv5

Class	mAP@50	mAP@50-95
Automatic Rifle	0.65	0.53
Bazooka	0.70	0.37
Handgun	0.90	0.73
Shotgun	0.90	0.66

SSD 300 ranks third, with a mAP@50 of 0.57 and mAP@50-95 of 0.36. Such low metrics indicate that the model struggles to accurately identify object boundaries. In addition to this issue, SSD 300 demonstrated low accuracy in correct predictions: specifically, a recall of 52% combined with a high number of false positives, as shown by a low precision of 36%. Likely, due to generating many predictions, including incorrect ones, the model captures some true positives, thereby increasing recall. However, this also means that the model did not effectively learn to recognise the desired objects in images.

RetinaNet showed the worst results. Its mAP@50 and mAP@50-95 metrics were just 0.38 and 0.20, respectively. This clearly indicates that the model does not effectively detect correct object boundaries. Precision was only 20% yet recall reached an unexpectedly high 48%. Based on these figures, many false positives were observed, reflecting low model selectivity. RetinaNet essentially failed to learn how to properly detect various types of weapons in images.

The results described above can be attributed, in part, to limitations in the dataset. For YOLOv5 and Faster R-CNN, dataset limitations were the main obstacle to achieving more accurate predictions. The dataset

has several systemic issues that collectively prevented these models from performing better. Specifically, the dataset contains 9 classes of various weapons but only 439 training images, which is insufficient to teach the model to correctly detect each class. Indirect confirmation of this issue is shown in Table 3 below.

As seen from the table, mAP@50 and mAP@50-95 scores differ significantly across weapon types, indicating that the model particularly struggled to identify automatic rifles and bazookas in images. At the same time, this issue was not due to an insufficient number of epochs. YOLOv5 and Faster R-CNN extracted the maximum amount of information available from the dataset used. This conclusion can be drawn from the graphs illustrating the growth of mAP@50 and mAP@50-95 over epochs.

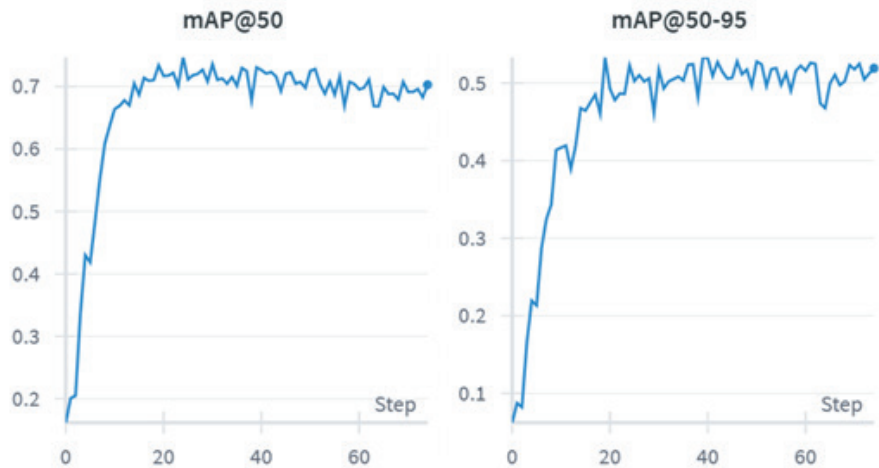


Figure 3. The evolution of accuracy (mAP) metrics on the validation set for Faster R-CNN

As evident from the graphs above, both metrics plateaued around the 20th epoch, indicating that further increases in epoch count would be pointless. Expanding the original dataset would yield a greater improvement.

Regarding SSD 300, the most probable reason for its poor results was the limitation of input image size. The dataset contains images of various dimensions, and resizing larger images down to 300x300 inevitably results in losing some low-level features. This leads to degradation in the results for mAP@50 and mAP@50-95 metrics for this model.

For RetinaNet, which demonstrated the lowest performance, various input resolutions were tested, but the most effective was found to be 512x512. Despite this, RetinaNet exhibited a similar plateau on the accuracy graph as shown in Figure 3, likely indicating that RetinaNet's internal algorithms are less effective compared to the other tested models [16].

4. CONCLUSION

This research addressed the core research question by comparatively evaluating four object detection models, YOLOv5, SSD 300, RetinaNet, and Faster R-CNN, on a security-relevant dataset. Despite certain systematic issues with the dataset, two models achieved good results according to the metrics mAP@50 and mAP@50-95. YOLOv5 demonstrated the best performance across all parameters. The mAP@50 for this model reached 0.80, which is a respectable result, though it could be further improved by expanding and refining the dataset.

This work contributes to a clearer understanding of the strengths and limitations of these popular object detection models in the context of weapon detection, potentially assisting researchers and industry practitioners in selecting and adapting models for real-time security applications. Additionally, the study highlighted potential shortcomings in SSD 300 and RetinaNet, offering insights for future improvements.

Future research should focus on evaluating these models on more diverse and larger datasets, exploring fine-tuning strategies, and testing in real-world surveillance scenarios to assess their robustness and generalisation capabilities.

REFERENCES

- [1] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using Deep Convolutional Neural Network Architectures for Object Classification and Detection within X-ray Baggage Security Imagery", Accessed: Mar. 30, 2025. [Online]. Available: <https://durham-repository.worktribe.com/output/1338347/>
- [2] A. Egiazarov, F. M. Zennaro, and V. Mavroedis, "Firearm Detection via Convolutional Neural Networks: Comparing a Semantic Segmentation Model Against End-to-End Solutions," Dec. 17, 2020, *arXiv: arXiv:2012.09662*. doi: 10.48550/arXiv.2012.09662.
- [3] G. Batsis, I. Mademlis, and G. T. Papadopoulos, "Illicit item detection in X-ray images for security applications," May 03, 2023, *arXiv: arXiv:2305.01936*. doi: 10.48550/arXiv.2305.01936.



- [4] K. J. Liang et al., "Toward Automatic Threat Recognition for Airport X-ray Baggage Screening with Deep Convolutional Object Detection," Dec. 13, 2019, *arXiv*: arXiv:1912.06329. doi: 10.48550/arXiv.1912.06329.
- [5] S. Yellapragada et al., "CCTV-Gun: Benchmarking Handgun Detection in CCTV Images," Jul. 11, 2023, *arXiv*: arXiv:2303.10703. doi: 10.48550/arXiv.2303.10703.
- [6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. CVPR*, May 2016, pp. 779–788. doi: 10.48550/arXiv.1506.02640.
- [7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jan. 06, 2016, *arXiv*: arXiv:1506.01497. doi: 10.48550/arXiv.1506.01497.
- [8] W. Liu et al., "SSD: Single Shot MultiBox Detector," vol. 9905, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.
- [9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," Feb. 07, 2018, *arXiv*: arXiv:1708.02002. doi: 10.48550/arXiv.1708.02002.
- [10] "(PDF) The Pascal Visual Object Classes (VOC) challenge." Accessed: Mar. 27, 2025. [Online]. Available: https://www.researchgate.net/publication/220659463_The_Pascal_Visual_Object_Classes_VOC_challenge
- [11] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," Feb. 21, 2015, *arXiv*: arXiv:1405.0312. doi: 10.48550/arXiv.1405.0312.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," vol. 8691, 2014, pp. 346–361. doi: 10.1007/978-3-319-10578-9_23.
- [14] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," Sep. 18, 2018, *arXiv*: arXiv:1803.01534. doi: 10.48550/arXiv.1803.01534.
- [15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.
- [16] Z.-Q. Zhao, P. Zheng, S. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," Apr. 16, 2019, *arXiv*: arXiv:1807.05511. doi: 10.48550/arXiv.1807.05511.



STUDENT SESSION

HANDWRITTEN DIGIT RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS AND BIG DATA PROCESSING

Pavle Matijašević*,
[0009-0007-4647-2555]

Miloš Mravik
[0000-0001-5442-3998]

Singidunum University,
Belgrade, Serbia

Abstract:

In this paper, we present a handwriting recognition system that combines neural networks and big data processing. We use a custom convolutional model, and optimized data augmentation techniques. To manage and process large volumes of data more efficiently, we relied on Apache Spark. Additionally, a user-friendly API has been created to enable real-time recognition of handwritten digits. Evaluation of the system on a custom dataset shows extremely high accuracy, with precision greater than 98% on the test data.

The main challenge now is to process large datasets, as well as manage different handwriting styles and ensure that the models perform accurately in real-world scenarios. The main contributions of this work are implementations of an efficient convolutional model for recognizing handwritten digits, a system for processing big data in a distributed environment, web API and user interface for real-time handwriting recognition, model error analysis and suggestions for further improvements.

Keywords:

Machine Learning, Big Data, Digit Recognition, Artificial Intelligence.

INTRODUCTION

Handwritten digit and symbol recognition is one of the most important areas in image processing and artificial intelligence itself, with wide applications in the areas of document digitization, automatic test scoring, and signature authentication. Traditional methods such as optical character recognition and hidden Markov models have shown limitations in recognizing handwritten digits and symbols due to large variability in writing styles. Handwritten digit recognition remains a challenge due to variations in handwriting, ambiguous symbols, and poor scan quality, and even a 1% recognition error may lead to serious issues in critical industries such as banking or healthcare. [1] With the development of deep neural networks, and with an emphasis on convolutional neural networks, we have seen significant performance improvements in handwritten digit recognition problems. Convolutional models can automatically recognize key features of handwritten characters, achieving high recognition accuracy. CNN architectures such as LeNet-5 have achieved state-of-the-art performance on digit tasks, with error rates as low as 0.7% on the MNIST dataset when augmented data is used. [2]

Correspondence:

Pavle Matijašević

e-mail:

matijasevic.pavle99@gmail.com





A breakthrough in the use of CNNs for large-scale image classification was achieved with the introduction of AlexNet [3], a deep CNN trained on the ImageNet dataset with over 1.2 million labelled images. This model drastically reduced the top 5 error rate from 26.2% (previous best) to 15.3%, demonstrating the power of deep architectures in visual recognition.

2. METHODOLOGY AND SYSTEM IMPLEMENTATION

2.1. GENERAL SYSTEM OVERVIEW

The handwriting recognition system consists of four main components:

- Collection and augmentation: Generating a large dataset of handwritten digits using augmentation techniques.
- Training Convolutional Neural Networks (CNN): Developing and Optimizing a Model for Handwritten digit recognition.
- Distributed data processing using Apache Spark: Efficient processing of large datasets before model training.
- Real-time handwriting recognition API: Fast API implementation is a service that allows users to recognize digits in real time via a web interface.

2.2. DATA PREPARATION AND PROCESSING

To train the model, we created a custom dataset of handwritten digits that is like the MNIST dataset but has additional variations in handwriting. The dataset was manually generated, and further augmented with augmentation techniques:

- Rotation (15, 7, 5, 20, 10)
- Scaling (enlarging and reducing size)
- Translation in arbitrary directions
- Adding noise to simulate handwriting variability
- Blurring to simulate poorly scanned images
- Changing brightness and contrast

Augmentation was performed using the Pillow library, while big data processing was done through Apache Spark for scalable processing.

2.3. THE ARCHITECTURE OF THE CNN MODEL

To recognize handwritten digits, we used a convolutional neural network (CNN) with the following architecture:

- Input layer: 28x28 pixel grayscale image.
- Convolutional layers: Two convolutional layers with a ReLU activation function
- Pooling layer: Maximum pooling (2x2) for dimensionality reduction.
- Fully connected layers: Two-layer with Softmax output for digit classification
- Dropout layers: Regularization of the model to reduce overfitting.

The network incorporated several innovations, including the use of ReLU activations, dropout to reduce overfitting, overlapping pooling, and parallel training on GPUs. These architectural and training choices made it possible to successfully train a network with 60 million parameters and achieve state-of-the-art results on complex visual tasks [3]. PyTorch was used to train the model, optimization was performed using the Adam optimizer, and the loss function was Cross-Entropy Loss.

2.4. DISTRIBUTED PROCESSING USING APACHE SPARK

Apache Spark was used to process large datasets before training the model, which allows for: Efficient loading and filtering of data from a large set of images, parallel augmentation using Spark RDD and PySpark functions, and faster data preprocessing before training the model.

This approach enables scalable management of large data sets and reduced processing time.

2.5. REAL-TIME HANDWRITING RECOGNITION API

The system allows users to recognize handwritten digits in real time via Fast API service that:

- Receives an image of handwritten digit via a web application.
- Converts the image to a format suitable for the CNN model.
- Runs a prediction using the trained CNN model and returns the result
- Allow visualization of predictions on the frontend application.



The user interface is implemented using Vue.js and Bootstrap, while the API allows for the fast and accurate request processing.

2.6. MODEL PERFORMANCE EVALUATION

The trained model was tested on a test dataset, with the key metrics being:

- Model accuracy
- Precision and responsiveness have high values on all digit classes
- Error analysis

3. EXPERIMENTS AND EVALUATION OF RESULTS

For training and learning the model, we used a dataset of 100,000 images of handwritten digits, including original and augmented images. To assess the quality of the model, we used standard classification metrics such as accuracy, precision, response, f1 value, and confusion matrix.

Similar evaluations conducted in recent studies have confirmed that CNN models consistently outperform other machine learning methods in terms of accuracy, precision, and robustness. In a 2024 study, the CNN model achieved 99.31% test accuracy, surpassing MLP and SVM models, which scored 96.89% and 95.31%, respectively. [1]

Table 1 shows the results of our CNN model on the test set:

Table 1. Results of the CNN model on the test set

	Precision	Recall	F1-score
0	1.0	1.0	1.0
1	0.89	1.0	0.94
2	1.0	0.9	0.95
3	0.95	1.0	0.98
4	1.0	1.0	1.0
5	0.89	1.0	0.94
6	1.0	1.0	1.0
7	1.0	1.0	1.0
8	1.0	1.0	1.0
9	1.0	0.84	0.91
Accuracy			0.98
Macro average	0.97	0.97	0.97
Weighted average	0.98	0.98	0.98

Figure 1 shows an example of a successfully classified digit:

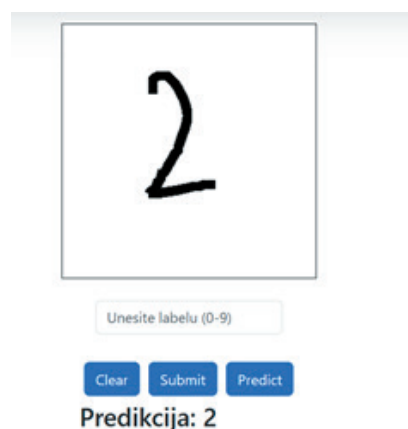


Figure 1. Successful classified digit



In Figure 2, we show example of unsuccessfully classified digit:



Figure 2. Unsuccessful classified digits

In Table 2, we show a comparison of our model with classical OCR methods and standard deep learning models:

Table 2. Model comparison results

Column 1	Column 2
Model	Accuracy
Tesseract OCR	86.5%
LaNet 5	97%
Our model	98.2%

By analysing the incorrect predictions, we identified several examples where the model had problems. Generally, the model fails in the following cases: Badly written digits, overlap between numbers, and too dark or bright images.

These results are consistent with previous studies, where LaNet-5 also outperformed classical methods, achieving 99.3% accuracy when trained with augmented data [2].

Tesseract OCR, although widely used due to its open-source nature and historical significance, achieves a significantly lower accuracy compared to CNN-based methods. The original Tesseract v2.0 achieved character-level error rates ranging from 1.61% to 2.22% on standard test sets, which is notably worse than CNN-based models like LeNet-5 or the proposed system in this work [4].

Techniques such as ridge regression-based feature selection have also shown great potential in predictive modelling tasks by improving generalization and reducing overfitting. [5] In addition, CNN-based systems have demonstrated remarkable success in healthcare applications, where accurate recognition of medical patterns is critical. [6]

The achieved model accuracy is 98.2%, which far surpasses classic OCR system like Tesseract. The model was trained on a dataset of 100,000 images, including augmented versions, which allowed it to better generalize different users handwriting. The most common errors occurred with visually similar digits (3-5, 1-7), while the other digits had the lowest error rates. Apache Spark significantly accelerated data processing and enabled efficient management of large datasets. Fast API enabled fast image processing and real-time prediction retrieval.

Although our model achieves high accuracy, we have identified certain challenges that may be the subject of future research:

- Visual similarity of digits: Some digits are very similar to each other (3 and 5), which can cause errors even in more advanced CNN models.
- Different handwriting styles: The model is trained on a set of handwritten digits but may also have difficulty with highly stylized or illegible digits.
- Extreme lighting conditions: Images that are too bright or dark can make recognition difficult, despite the normalization techniques applied.



Future research will focus on extending the dataset to include letters and symbols, using advanced neural networks such as the Transformer architecture for improved analysis of handwriting sequences, as well as combining convolutional neural networks (CNNs) with recurrent neural networks (RNNs). Additionally, efforts will be directed toward real-time optimizations and the development of a mobile application. In addition to Transformer-based models, deep residual networks offer a powerful and scalable approach to improving handwriting recognition. ResNets alleviate the degradation problem that occurs with increasing network depth by introducing identity shortcut connections, enabling the training of models with over 100 layers and achieving a top 5 error of 3.57% on the ImageNet test set [7].

4. CONCLUSION

The model based on convolutional neural networks achieved very good results and proved to be reliable in practical use. Most of the errors happened when the digits looked like each other, for example, 3 and 5, or 1 and 7. This setup also includes an API that responds in real time, and several data augmentation methods were applied to help the model perform better and generalize across different handwriting styles.

Similar techniques that focus on generalization and robustness, such as regularized feature selection, have shown success in reducing overfitting and improving model performance, especially when compared to traditional approaches like SVM or kNN.

Moreover, architectures like ours have been successfully applied in real-world domains, including the medical field, where convolutional models are used for precise pattern recognition and decision support based on visual data.

All of this shows that combining deep learning with scalable processing tools can produce fast and accurate systems, even for tasks that involve a lot of variation, like recognizing handwritten digits.

REFERENCES

- [1] N. F. A. M. Raheem, "Handwritten digit recognition using machine learning," *Journal of Theoretical and Applied Information Technology*, 2024.
- [2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol.86, no. 11, pp. 2278-2324, 1998.
- [3] A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.
- [4] R. Smith, "An Overview of the Tesseract OCR Engine," in *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, 2007.
- [5] M. Mravik, T. Vetriselvi, K. Vankatachalam, M. Sarac, N. Bacanin, S. Adamovic, "Diabetes Prediction Algorithm Using Recursive Ridge Regression L2," *Computers, Materials & Continua*, 2022.
- [6] M. Sarac, M. Mravik, D. Jovanovic, I. Strumberger, M. Zivkovic, N. Bacanin, "Intelligent diagnosis of coronavirus with computed tomography images using a deep learning model," *Journal of Electronic Imaging*, vol. 32, no. 2, 2022.
- [7] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.



SDLC-INDEPENDENT PYTHON-BASED QUERY PERFORMANCE BENCHMARKING APPROACH AND PRACTICAL OPTIMAL DATABASE SELECTION GUIDELINES

Katarina Milojković*,
[0000-0001-8658-8973]

Petar Spalević,
[0000-0002-6867-7259]

Nikola Vasić,
[0000-0003-2713-4589]

Nikola Milojković,
[0009-0001-0106-9662]

Hristina Milojković
[0000-0003-4150-3301]

Singidunum University,
Belgrade, Serbia

Abstract:

One of the critical decisions of software development teams in application development is choosing the most optimal database. Modern business conditions require software development teams to continually improve application performance, and a common dilemma is whether to transition from a relational database to a non-relational database or vice versa. Changes during the application development phase can slow down and complicate the process, so it is crucial to empirically approach data analysis and decision-making before implementing any changes. This research aims to facilitate the optimal database selection by following established practical guidelines for optimal database selection and implementing an SDLC-independent Python-based query performance benchmarking approach. This benchmarking approach is a crucial part of the optimal database selection process particularly useful in the early stages of development or when considering a migration to an existing project. The research methodology includes qualitative and quantitative methods: analytical-synthetic, experimental, comparative analysis, and hypothetical-deductive methods. The results of this research include the practical optimal database selection guidelines, the process of conducting the benchmark, and the utilization of both the guidelines and benchmark results for optimal database selection of an application where changing the initially selected MySQL database to MongoDB is being considered.

Keywords:

SQL vs NoSQL, MySQL vs MongoDB, Optimal Database Selection Guidelines, Query Performance Benchmark, Python.

INTRODUCTION

In modern business conditions, software development teams frequently encounter challenges such as difficult application development, data consistency and integrity, processing complex queries, real-time fraud detection, robustness, efficiency, scalability, agility, transaction consistency, complex relationships and transactions, unstable workloads, real-time analytics and updates, managing user-generated data, and many more. The main reasons for those challenges can be wrong initial database selection, requirements evolving over time, increased application popularity, unpredictable traffic spikes, large or constantly evolving data sets, and changed market conditions and consumer preferences. In order to resolve those challenges, it is vital to select the most optimal database for the application's architecture and requirements.

Correspondence:

Katarina Milojković

e-mail:

katarina.milojkovic.21@singimail.rs



All of the above confirms the extreme importance of choosing the right database solution at every stage of the Software Development Life Cycle (SDLC). Developers can choose a relational (SQL) or a non-relational (NoSQL) database solution for the sake of establishing, maintaining, and improving performance, security, reliability, and user satisfaction. Changes during the application development phase can slow down and complicate the process, especially if that change is a wrong database selection because replacing it will be time-consuming and costly. In order to avoid that, it is crucial to have a deep understanding of SQL and NoSQL pros and cons, as well as empirically approach data analysis and decision-making before implementing any changes. Organizations conduct benchmarks using a framework or script created by a programming language such as Python in order to analyze the data flow and performance in real-time as well as simulate other database solutions working in an application and test their performance. The goal of this research is to share insights and knowledge regarding optimal database selection facilitation by following established practical guidelines for optimal database choice and implementing an SDLC-independent Python-based query performance benchmarking approach.

The hypothesis tested in this research states that MongoDB is the optimal database solution for the “BooksByHM” application. The content-sharing web application “BooksByHM” allows authors to publish chapters, books, images, and audio versions while enabling users to interact through likes, ratings, comments, and purchases. The application is currently in a mid-developing phase where changing the initially selected MySQL database to MongoDB is being considered.

In the ‘Literature review’ section, relevant research was summarized and presented to provide a deeper insight into this research. This section serves as a foundation for practical optimal database selection guidelines. The methodology used for this research is thoroughly explained in the ‘Methodology’ section, while the implementation of the Python-based query performance benchmarking approach is in the ‘Implementation’ section. The results of qualitative and quantitative methods used in this research and the discussion of those results are shown in the ‘Results and Discussion’ section, whereas the ‘Conclusion’ section presents the conclusions drawn from the research.

2. LITERATURE REVIEW

Database selection is one of the most crucial decisions developers must make to develop a system and maintain performance, reliability, and user satisfaction. [1] [2] Selecting the wrong database will make application development difficult because replacing it will be time-consuming and costly, so it must be done with careful consideration of long-term support and sustainability. [1] Before choosing, switching, or migrating databases, organizations conduct benchmarks using a framework or script created by a programming language such as Python in order to analyze the data flow and performance in real-time. [3] [4] [5] Benchmarking is the process of running a specific program or workload on a machine or system to evaluate its performance for that workload accurately. [6] The reason behind making database changes is to improve efficiency, maintainability, scalability, and security. [5] With the Python migration script, data from one database can be present in another database. [5] Python is a dynamically typed programming language frequently used for scientific research, web development, machine learning, artificial intelligence, and data analysis. [7] Python is usable in various fields because it has a powerful standard library and wide module support. [7] Working with MySQL and MongoDB is possible in Python if libraries such as mysql-connector-python and pymongo are installed and MySQL and MongoDB servers are connected. [8]

In order to choose the most optimal database, relational (SQL) or non-relational (NoSQL), several factors must be taken into account such as the data and database structure, schema flexibility, data scalability, query language for defining and manipulating the data, performance indicators, guaranteed properties (ACID, BASE, CAP), and cost. [9] Another important factor can be security, licensing, and its capability with different tools. [5] SQL offers extensive integration support, while NoSQL provides modern APIs and flexible data formats for seamless integration with microservices and cloud-native architectures. [2] These factors can be more harmful than helpful if not utilized efficiently within the software architecture [5] and requirements. [8]

Based on the data structure of stored data, database models are examined as relational databases (SQL databases), where structured data is stored in a predefined schema, and non-relational databases (NoSQL databases), where unstructured data is not stored in a predefined schema. [7] The advantages of storing data in a predefined schema are predicting entities and values the application expects, validating data based on exist-



ing or new records, using database constraints, and organizing data through normalization. [1] On the other hand, NoSQL schemas enhance flexibility and scalability, simplify Big Data management, and allow developers to focus on software application development, and database optimization. [9] SQL databases, as a tabular relational model, use tables where data is stored in the forms of rows (records) and columns (attributes). [10] [8] Popular SQL databases that have become industry standards are MySQL, PostgreSQL, Oracle Database, and Microsoft SQL Server. [11] NoSQL databases are non-tabular databases, which is why they use data models such as document-based, key-values, column-based, or graph-based. [3] Document-based databases store data in document forms inside a collection. [1] Since they do not implement field validations and constraint checks, they are swift. [1] Because the same entity with different fields can be stored multiple times, developers need to pay attention to the application side to avoid making mistakes. [1] Document-based databases have a multi-server architecture. [1] They are ideal for application development, flexible data structures, scalability, and dynamic schemas. [1] Popular document-based NoSQL solutions are MongoDB, CouchDB, CosmosDB, DynamoDB [1] [10], and Firebase. [12] MongoDB is an open-source NoSQL database that uses a document-oriented approach. [4] MongoDB uses XML, JavaScript Object Notation (JSON), or Binary JSON (BSON) to encode and store data in a document format. [10] Queries are performed over collections or using map-reduce. [4]

MySQL has a networked client-server architecture which has two components, MySQL server and client programs. [4] MySQL architecture is a web of task-related functions that work to finish the job of a database server. [4] Relational databases can have a multi-server architecture by using shared storage technology. [1] MySQL uses master-slave replication through a cluster-based architecture, whereas MongoDB uses master-slave replication through replica sets. [4]

Most relational databases are vertically scalable, meaning that the load on a server will be increased by increasing/upgrading the server's hardware components like RAM, HDD/SSD, and CPU. [1] Non-relational databases are horizontally scalable, meaning they can support increased traffic by adding servers and instances. [10] NoSQL databases can be more cost-effective than SQL databases due to horizontal scalability and open-source free solutions. [9] The scalability of MongoDB is easy to implement and performs better whereas MySQL maintains data integrity. [4] [13] [14] However, MySQL can improve scalability by using cloud-based technologies. [14]

SQL databases use Structured Query Language (SQL), a declarative and standardized query language for defining, manipulating, querying, and managing data. [9] [10] [11] NoSQL databases use Not Only Structured Query Language (NoSQL), a custom query language tailored to their specific data models. [3]

The performance evaluation (query runtime, memory used, CPU used, and storage size) of different SQL and NoSQL databases resulted in MongoDB outperforming almost all tests with a large data volume [13] [3] for example, 10,000+ records. [10] The result is due to the way data is stored, complex joins, and data normalization. [10] NoSQL select operations are 3 times faster, delete operations are 6 times faster, update operations are 9 times faster, and insert operations are 15 times faster than SQL. [10] MySQL shows better performance for small datasets (a few thousand records) and few database operations (a hundred operations daily). [15] SQL databases have a better join query performance [13] and use less CPU resources and memory usage for task completion compared to NoSQL databases. [15] However, the process of storing and retrieving complex data types (images as byte data) was faster inside NoSQL databases like MongoDB. [15] MongoDB outperformed MySQL in terms of Latency, Throughput, scalability, security, performance, and availability. [4]

SQL databases follow ACID (Atomic, Consistent, Isolated, and Durable) transaction principles [7] and CAP (Consistency, Availability, and Partition Tolerance) theorem. [11] Instead, NoSQL databases follow the BASE (Basically Available, Soft-State, and Eventual Consistency) transaction principles [7] and CAP (Consistency, Availability, and Partition Tolerance) theorem. [11] NoSQL databases often do not follow ACID principles, such as strong data consistency, which makes processing complex SQL queries challenging. [14] This can be overcome with automatic machine-learning classification techniques such as SVM, K-means, and NBC. [14] Distributed systems can only prioritize two out of three CAP principles. [11] SQL databases prioritize consistency and availability, whereas NoSQL databases prioritize tolerance and availability which makes them offer eventual consistency instead of strict consistency. [11] NoSQL databases lack full support for atomicity, consistency, isolation, and durability features found in SQL databases. [14] NoSQL databases sacrifice some robustness to achieve more speed and scalability. [1] Relational databases managed by RDBMS assure data integrity and transaction consistency. [9] In relational databases, the data storage performance degrades as the data volume increases. [9]



SQL databases are used for e-commerce [8], transaction applications, financial systems, enterprise resource planning (ERP) systems, customer relationship management (CRM) systems, applications that require complex queries [11], relationships and transactions, strong data integrity, inventory control, applications with stable workloads [2], applications with multi-row transactions [9], payment processing, and core banking systems. [16]

NoSQL databases are used for big data analytics, recommendation systems, Internet of Things (IoT) [8], cloud-based applications [4], real-time analytics, content management systems, social networks [11], applications with large or constantly evolving data sets [9], applications focused on rapid data access, flexible data models, scalability, high traffic loads, unpredictable traffic spikes, managing user-generated data, real-time updates, product catalogs, and inventory data [2], agile applications where requirements evolve over time [17] [11], data mining applications [13], real-time fraud detection, and personalized finance services. [16]

A hybrid database approach combines the strengths of SQL and NoSQL systems allowing organizations to tailor their database solutions to their needs. [11] [8] [2] However, this approach requires careful planning and implementation because it can complicate data management and integration. [11] A hybrid solution (SQL for transaction processing and core banking systems, and NoSQL for big data analytics, real-time fraud detection, and customer management) was chosen for the FinTech application founded on large-scale data processing, transactional integrity, and real-time analytics, and warrants robust and highly scalable database solutions. [16] Netflix uses SQL for billing and subscriber data while using NoSQL for viewing history and recommendations. Similarly, Uber uses SQL for transactional accuracy in rides and payments, while NoSQL handles real-time tracking for high availability. [11] PayPal uses Apache Cassandra to power its real-time fraud detection systems. [16] JP Morgan Chase uses SQL databases in its core banking operations. [16] Countries with a higher number of train stations and stops such as Germany, Netherlands, and others, use non-relation databases. Whereas smaller size train stations in Slovakia use relation databases. [10] Google, Facebook, Twitter, and Amazon prefer NoSQL database systems because they have very large datasets, and they need to implement their solutions on multiple servers and NoSQL are horizontally scalable databases. [1] [4] [10]

3. METHODOLOGY

The research combines qualitative and quantitative methods such as analytical-synthetic, experimental, comparative analysis, and hypothetical-deductive methods. The analytical-synthetic method was used to conduct a comprehensive analysis of SQL and NoSQL databases and synthesize those findings in the form of practical guidelines for choosing the optimal database. The experimental method was used to conduct an experiment with the SDLC-independent Python-based query performance benchmarking approach to measure the query execution time of MySQL and MongoDB under the same conditions. The benchmark results were compared and analyzed using the comparative analysis method. A hypothetical-deductive method was used for testing the set hypothesis with the results of analytical-synthetic, experimental, and comparative analysis methods, and reaching deductions on optimal database choice for the “BooksByHM” application based on the obtained results.

4. IMPLEMENTATION

The process of conducting an SDLC-independent Python-based benchmark for measuring query execution time consists of database and data preparation, setting up benchmarking conditions, executing benchmarking queries, and measuring performance metrics.

The experiment is set under the same initial benchmarking conditions:

- Hardware and system specifications:
 - Processor: AMD Ryzen 7 3700U with Radeon Vega Mobile Gfx (8 CPUs), ~2.3GHz
 - RAM: 8.00 GB
 - Operating System: Windows 11 Pro
 - System type: 64-bit operating system, x64-based processor
- Software and database versions:
 - Python version: 3.13.2
 - PIP version: 24.3.1
 - MySQL version: 10.4.28-MariaDB
 - MongoDB version: 8.0.0
 - Python libraries: mysql-connector-python, pymongo, random, and time.



- Dataset description:
 - The smaller dataset contains 10 book records and 11 user records.
 - The bigger dataset contains 10.000 book records and 1.001 user records.
 - MySQL tables: books, users, user_books (many-to-many relationship)
 - MongoDB collections: books, users (embedded book list)
 - MySQL indexing: primary keys on id fields, foreign keys in user_books
 - MongoDB indexing: default index on _id.

The databases were prepared by duplicating the existing MySQL database used in the “BooksByHM” application and creating an equivalent MongoDB database. Working with MySQL and MongoDB databases in Python was made possible by installing mysql-connector-python and pymongo Python packages and connecting MySQL and MongoDB servers in the Python script. Installing the MySQL Connector package for Python on Windows using PIP was done with the command: *pip install mysql-connector-python*. On the other hand, the command used for installing the PyMongo package for Python on Windows using PIP was: *pip install pymongo*. Listing 1 shows the Python code used to establish the connection of specific MySQL and MongoDB databases.

```
mysql_conn = mysql.connector.connect(
    host="localhost",
    user="root",
    password="",
    database="books_by_hm"
)
mysql_cursor = mysql_conn.cursor()

mongo_client = MongoClient("mongodb://localhost:27017/")
mongo_db = mongo_client["books_by_hm"]
mongo_books = mongo_db["books"]
mongo_users = mongo_db["users"]
```

Listing 1. Python code is used to establish a connection between the MySQL database and the MongoDB database

```
def measure_mysql():
    start = time.time()
    mysql_cursor.execute("SELECT * FROM books WHERE name='Book5000'")
    mysql_cursor.fetchall()
    end = time.time()
    print(f"MySQL - simple query: {end - start:.6f} seconds")

    start = time.time()
    mysql_cursor.execute("""
        SELECT users.*
        FROM users
        JOIN user_books ON users.id = user_books.user_id
        JOIN books ON books.id = user_books.book_id
        WHERE books.name = 'Book5000'
    """)
    mysql_cursor.fetchall()
    end = time.time()
    print(f"MySQL - complex query: {end - start:.6f} seconds")

def measure_mongodb():
    start = time.time()
    mongo_books.find({"name": "Book5000"})
    end = time.time()
    print(f"MongoDB - simple query: {end - start:.6f} seconds")

    start = time.time()
    mongo_users.find({"books": {"$elemMatch": {"name": "Book5000"}}})
    end = time.time()
    print(f"MongoDB - complex query: {end - start:.6f} seconds")
```

Listing 2. measure_mysql() and measure_mongodb() functions written in Python



Tables and collections were created in the Python script. Also, the data used for the benchmark was prepared by automatically inserting the small and large datasets in the Python script.

Based on the application's needs, the chosen simple benchmarking query was finding a book by title, and the complex benchmarking query was finding users who own a specific book. Execution time in seconds was measured using Python's `time.time()` function before and after query execution. The scalability tests were performed by scaling datasets from 10 to 10,000 books and 11 to 1,001 users to analyze the performance fluctuation. Listing 2 shows `measure_mysql()` and `measure_mongodb()` functions written in Python used to execute queries and measure execution time.

5. RESULTS AND DISCUSSION

The measured query execution time of MySQL and MongoDB with different dataset sizes are shown in Table 1.

As shown in Table 1, MySQL simple query execution time is 11.9864 times faster for smaller datasets and 3.2198 times faster for bigger datasets than MongoDB. The result is due to the way data is stored, database constraints, and data normalization. MySQL is faster than MongoDB because it uses an index and searches directly on the table. As the data volumes increase, query execution time increases for both databases with MySQL remaining to be faster. This indicates that indexed search in MySQL scales is better for simple queries than in MongoDB.

Table 1. Benchmark results

Dataset size	MySQL		MongoDB	
	Smaller dataset	Bigger dataset	Smaller dataset	Bigger dataset
Simple query execution time (seconds)	0.000516	0.003316	0.006185	0.010677
Complex query execution time (seconds)	0.003165	0.011179	0.000073	0.000045

Table 2. Practical guidelines for optimal database selection

SQL database characteristics	tabular data models for storing structured data in a fixed schema; vertical scalability (increasing/upgrading the server's hardware components); scalability can be improved by using cloud-based technologies; client-server architecture; can have a multi-server architecture by using shared storage technology; using Structured Query Language (SQL) for defining, manipulating, querying, and managing data; better performance for small/limited datasets and few database operations; better join query performance; less CPU resources and memory usage for task completion; master-slave replication through a cluster-based architecture; ACID database transaction model; follows CAP theorem (prioritizing consistency and availability); strong data consistency; assure data integrity and transaction consistency; the data storage performance degrades as the data volume increases; offers extensive integration support;
NoSQL databases characteristics	non-tabular data models such as document-based, key-values, column-based, or graph-based for storing unstructured data in a flexible schema; cost-effective due to horizontally scalable (increasing/upgrading servers and instances) and open-source free solutions; scalability is easy to implement; multi-server architecture; using Not Only Structured Query Language (NoSQL) tailored to their specific data models for defining, manipulating, querying, and managing data; better performance for large datasets; faster storing and retrieving complex data types (images as byte data); master-slave replication through replica sets; BASE database transaction model; follows CAP theorem (prioritizing tolerance and availability); eventual data consistency; lack full support for ACID features; less robustness; high-velocity and scalability; provides modern APIs and flexible data formats for seamless integration with microservices and cloud-native architectures;
SQL is suitable for	e-commerce applications; transaction applications; financial systems; enterprise resource planning (ERP) systems; customer relationship management (CRM) systems; applications that require complex queries, relationships, and transactions; applications that require strong data integrity; inventory control; applications with stable workloads; applications with multi-row transactions; payment processing; core banking systems;
NoSQL is suitable for	big data analytics; recommendation systems; Internet of Things (IoT); cloud-based applications; real-time analytics; content management systems; social networks; applications with large or constantly evolving data sets; applications focused on rapid data access, flexible data models, and scalability; high traffic loads; unpredictable traffic spikes; managing user-generated data, real-time updates, product catalogs, and inventory data; agile applications where requirements evolve over time; data mining applications; real-time fraud detection; personalized finance services;



As shown in Table 1, MongoDB complex query execution time is 43.3562 times faster for smaller datasets and 248.4222 times faster for bigger datasets than MySQL. The result is due to the way data is stored, complex join, and data normalization. MongoDB is significantly faster than MySQL because the search is performed directly, while MySQL requires joining data from multiple tables and filtering the results. When the database is large, MongoDB remains consistently fast (even faster than with a smaller database), while MySQL's execution time increases significantly. This suggests that MongoDB is better for complex queries as it does not need expensive JOIN operations like MySQL. As data grows, MySQL's complex queries become significantly slower, while MongoDB remains efficient.

Utilizing both the guidelines, shown in Table 2, and benchmark results, shown in Table 1, the most optimal database solution for the 'BooksByHM' application is a hybrid database approach that combines the strengths of MySQL and MongoDB. For optimal performance, MySQL would be used for structured data (users, purchases, and ratings), transaction processing, core banking operations, and dynamic content generation (SEO optimization). On the other hand, MongoDB would be used for content storage and interactions (books, comments, and likes), real-time analytics, real-time fraud detection, customer management, large-scale data processing, personalized recommendations, storing and retrieving images (as binary data) and audiobook files, searching books, managing nested comments, and generating statistical insights.

6. CONCLUSION

The research contributes by presenting practical guidelines for optimal database selection and an SDLC-independent Python-based query performance benchmark. The hypothesis was successfully tested with both the guidelines and benchmark results. The hybrid approach can utilize the strengths of both MySQL and MongoDB by allowing tailored database solutions to the "BooksByHM" application's needs. The benchmarking approach highlighted in this research can be used to analyze the data flow and performance in real-time, as well as simulate database solutions and test their performance without needing to have a concrete application. This is the reason why this benchmarking approach is SDLC-independent. Following the established practical guidelines for optimal database selection and imple-

menting the SDLC-independent Python-based query performance benchmarking approach is particularly useful in the early stages of development or when considering a migration to an existing project. The research highlights the importance of selecting the right database solution and demonstrates that the benchmarking process can be both simple and efficient using Python. It aims to encourage and motivate developers to experiment with and test different database solutions. Additionally, it serves as an accessible and engaging experiment that anyone can try in their spare time. However, this study represents just the tip of the iceberg in the field of benchmarking. Future directions of this research could explore additional benchmarking methods using Python to further enhance database performance evaluation.

REFERENCES

- [1] H. Paci, "SQL vs NoSQL databases from developer point of view," *Industry 4.0*, vol. VII, no. 3, pp. 95-97, 2022.
- [2] T. Ramzan and G. Alwin, "Comparative Study of SQL vs. NoSQL for High-Performance E-commerce Databases," pp. 1-18, 2023.
- [3] M. Z. Khan, F. U. Zaman, M. Adnan, A. Imroz, M. A. Rauf and Z. Phul, "Comparative case study: An evaluation of performance computation between SQL and NoSQL database," *Journal of Software Engineering*, vol. I, no. 2, pp. 14-23, 2023.
- [4] R. Pandey, "Performance benchmarking and comparison of cloud-based databases MongoDB (NoSQL) vs MySQL (Relational) using YCSB," *Nat. College Ireland, Dublin, Ireland, Tech. Rep*, 2020.
- [5] D. Liberman, "Migration from NoSQL to SQL," 2023.
- [6] R. H. Saavedra and A. J. Smith, "Analysis of benchmark characteristics and benchmark performance prediction," *ACM Transactions on Computer Systems (TOCS)*, vol. XIV, no. 4, pp. 344-384, 1996.
- [7] M. Yeşilyurt and Y. Z. Ayik, "Comparison of C# and Python programming languages in terms of performance and coding on SQL server DML operations," *NanoEra*, vol. IV, no. 1, pp. 23-33, 2024.
- [8] Є. С. Тимошенко, "Робота з базами даних в python: SQL та NOSQL," 2024.
- [9] A.-G. Babucea, "SQL OR NoSQL DATABASES? CRITICAL DIFFERENCES," *Annals of Constantin Brancusi University of Targu-Jiu. Economy Series/ Analele Universității Constantin Brâncuși din Târgu-Jiu Seria Economie*, no. 1, 2021.



- [10] R. Čerešňák and M. Kvet, "Comparison of query performance in relational a non-relation databases," *Transportation Research Procedia*, vol. XL, pp. 170-177, 2019.
- [11] Y. Jani, "The role of sql and nosql databases in modern data architectures," *International Journal of Core Engineering & Management*, vol. VI, no. 12, pp. 61-67, 2021.
- [12] K. Milojković, M. Živković and N. B. Džakula, "Agile Multi-user Android Application Development With Firebase: Authentication, Authorization, and Profile Management," *Sinteza 2024-International Scientific Conference on Information Technology, Computer Science, and Data Science*. Singidunum University, pp. 405-412, 2024.
- [13] J. Antas, R. R. Silva and J. Bernardino, "Assessment of SQL and NoSQL systems to store and mine COVID-19 data," *Computers*, vol. XI, no. 2, p. 29, 2022.
- [14] R. A. Kadir, E. S. M. Surin and M. R. Sarker, "A Systematic Review of Automated Classification for Simple and Complex Query SQL on NoSQL Database," *Computer Systems Science & Engineering*, vol. XLVIII, no. 6, 2024.
- [15] S. A. FADHEL and E. A. JAMEEL, "A Comparison between NoSql and RDBMS: Storage and Retrieval," *International Journal of Applied Sciences and Technology*, vol. IV, no. 3, pp. 173-184, 2022.
- [16] P. Gowda and A. N. Gowda, "SQL vs. NoSQL databases: Choosing the right option for FinTech," *Journal of Scientific and Engineering Research*, vol. VII, no. 8, pp. 100-104, 2020.
- [17] D. Milojković and K. Milojković, "Improving the Business Resilience of an Organization by Applying Agile Project Managemet Approach," *FINIZ 2022-Business Resilience in a Changing World*, pp. 98-103, 2022.



MULTIMODAL RETRIEVAL-AUGMENTED GENERATION IN KNOWLEDGE SYSTEMS: A FRAMEWORK FOR ENHANCED SEMANTIC SEARCH AND RESPONSE ACCURACY

Marko Mihajlović*
[0009-0006-7979-0341]

Singidunum University,
Belgrade, Serbia

Abstract:

Maintaining structured and up-to-date documentation remains a critical challenge in knowledge management, especially as organizational data becomes increasingly unstructured and multimodal. This paper presents a framework for a Multimodal Retrieval-Augmented Generation (RAG) Knowledge Database Assistant, designed to enhance semantic search and improve the accuracy of generated responses. By combining retrieval-augmented generation techniques with support for diverse data modalities (e.g., text, images, and structured metadata), the proposed system mitigates hallucination risks and increases the reliability of information access. The framework enables precise, context-aware question answering, even when underlying knowledge repositories are incomplete or inconsistently maintained. Our approach demonstrates how multimodal integration and RAG pipelines can form a robust foundation for next-generation knowledge systems.

Keywords:

Multimodal Retrieval, Retrieval-Augmented Generation, Knowledge Management, Semantic Search, Question Answering.

INTRODUCTION

In modern knowledge management, organizations increasingly face the challenge of maintaining structured, accurate, and up-to-date documentation. While unstructured knowledge bases often contain a wealth of valuable information, their lack of formal organization poses significant obstacles to effective search, retrieval, and usage [1]. As the scale and complexity of digital knowledge assets continue to grow, traditional keyword-based search methods and manually curated documentation are no longer sufficient to meet users' expectations for precision and relevance [2].

To address these challenges, this paper introduces a Multimodal Retrieval-Augmented Generation (RAG) Knowledge Database Assistant, a framework designed to enhance the usability and searchability of unstructured and semi-structured information repositories. The system integrates advanced retrieval-augmented generation techniques with multimodal data processing to support both text and image-based queries, enabling accurate, context-aware answer generation with traceable references [3].

Correspondence:

Marko Mihajlović

e-mail:

marko.mihajlovic.23@singimail.rs



The proposed framework begins with exporting unstructured data into a defined format, which is then parsed using a custom-built tool to produce a structured XML representation of articles. Each article is automatically enriched with a synthetic question, generated by a language model (e.g., ChatGPT), and paired with a corresponding answer extracted from its content. These question-answer pairs, along with article titles and content, are embedded using OpenAI's Ada model, with weighted emphasis placed on the title and generated questions to optimize semantic representation.

The resulting embeddings are stored in a Redis database and indexed via RediSearch for efficient similarity-based retrieval [4]. During inference, a user's query is embedded and compared against the stored vectors to retrieve the top-matching articles. These are then combined with the original query and fed into a generative language model, which produces a coherent, accurate response enriched with references to the source materials. In addition to text-based interactions, the system also supports visual search. Image embeddings are created using the Vision Transformer (ViT) model and stored alongside textual metadata in the same vector database. Users can upload images or screenshots to retrieve related content, enabling a robust multimodal search experience that bridges textual and visual information.

By unifying structured parsing, semantic embedding, generative language modeling, and multimodal retrieval, the proposed framework delivers a scalable and extensible solution for navigating large, unstructured knowledge systems. It ensures high retrieval accuracy, minimizes hallucinated outputs, and supports a wide range of use cases where traditional search systems fall short.

2. METHODS

The following section outlines the design and implementation details of the proposed multimodal RAG-based framework.

2.1. DATA PREPARATION

The data preparation phase is responsible for transforming unstructured content into structured, semantically rich representations suitable for vector-based retrieval. This includes processing both textual and image-based data.

Articles are first extracted from unstructured sources (e.g., documentation, wikis) and structured into fields such as title, content, and a generated question. The questions are automatically created using a Large Language Model (LLM), such as ChatGPT, to semantically enrich the document and support better query matching. To embed the textual data, OpenAI's text-embedding-ada-002 model is used to generate 1536-dimensional vector representations for each article [5]. Fields are assigned weights to improve search relevance, titles and questions are prioritized (weight = 3.0), while content receives a lower weight (weight = 1.0) to minimize noise. These weights were empirically selected based on testing, where prioritizing the title and generated question over the full content yielded more accurate retrieval performance.

For visual data, the system extracts image references from structured XML. Using the OpenCLIP implementation of the ViT-B/32-quickgelu model, each image is preprocessed and encoded into a 512-dimensional normalized embedding vector [6]. Embeddings are cached locally to avoid redundant computation and are associated with metadata such as the image filename and post ID for later retrieval.

The framework utilizes two separate embedding pipelines, one for text and one for image data, each indexed in its own Redis vector database using RediSearch. This preparation pipeline ensures that both modalities, text and image, are encoded into high-quality embeddings, stored alongside useful metadata, and ready for indexing in the vector database. Both embedding pipelines leverage HNSW (Hierarchical Navigable Small World) indexing for efficient Approximate Nearest Neighbor (ANN) search, and cosine similarity is used to measure semantic proximity.

2.2. VECTOR DATABASE CONSTRUCTION

To enable scalable and efficient semantic retrieval, the system employs Redis as the underlying in-memory data store and RediSearch as the indexing and query engine. Redis is a high-performance, open-source, in-memory key-value store widely used for real-time applications due to its low-latency read/write operations. RediSearch is a Redis module that adds full-text search and secondary indexing capabilities, enabling complex queries, ranked retrieval, and schema definitions over Redis hashes [7].



In this framework, Redisearch is extended to support vector similarity search, transforming Redis into a vector database, a specialized type of database designed to store and retrieve high-dimensional vector representations (embeddings). Unlike traditional databases that rely on exact matching, vector databases enable approximate nearest neighbor (ANN) search, which is essential for retrieving semantically similar documents based on learned representations from models like OpenAI's Ada or OpenCLIP.

The system maintains two independent vector indexes, one for textual embeddings and another for image embeddings. Each index is tuned to the dimensionality and structure of its respective modality: 1536 dimensions for text (from Ada) and 512 for images (from ViT-B/32). These indexes are defined using Redisearch's schema configuration as shown in Listing 1 and leverage the HNSW (Hierarchical Navigable Small World) algorithm for fast and scalable ANN retrieval using cosine similarity as the distance metric. This setup allows the system to return semantically relevant results in real time, even across large datasets.

Text embeddings are stored with relevant metadata fields and configured weights to boost semantic precision during the search. Redis uses the HNSW indexing algorithm and cosine similarity for approximate nearest-neighbor retrieval.

Image embeddings are generated using OpenCLIP's ViT-B/32-quickgelu model, producing 512-dimensional normalized vectors. In addition to the embedding, each image post includes semantic metadata: title, content, and a permalink. These fields are stored in Redis using a hash and indexed with configurable weights as shown in Listing 2. This allows users to perform semantic searches using not only visual similarity, but also textual metadata tied to images (e.g., captions or context).

By separating these two vector spaces, the system supports modality-specific optimization and maintains retrieval precision across both text and image inputs. This modular structure is extensible and forms the foundation for building robust multimodal retrieval-augmented generation systems.

2.3. QUERY HANDLING AND ANSWER GENERATION

Once embeddings and metadata are stored in Redis, the system is equipped to handle real-time user queries and generate accurate, context-aware responses. The query processing pipeline supports both textual and visual inputs, enabling a flexible multimodal search experience.

When a user submits a text-based query, the system uses OpenAI's text-embedding-ada-002 model to generate a 1536-dimensional embedding of the query.

```
DEFINE TEXT_INDEX ON HASHES
PREFIX: "article_text:"
SCHEMA FIELDS:
- title: TEXT, WEIGHT = 3.0
- question: TEXT, WEIGHT = 3.0
- content: TEXT, WEIGHT = 1.0
- permalink: TEXT
- embedding: VECTOR
  TYPE = FLOAT32
  DIM = 1536
  INDEX_METHOD = HNSW
  DISTANCE_METRIC = COSINE
```

Listing 1. Text Index Schema Definition

```
DEFINE IMAGE_INDEX ON HASHES
PREFIX: "image_clip:"
SCHEMA FIELDS:
- title: TEXT, WEIGHT = 2.0
- content: TEXT, WEIGHT = 1.0
- permalink: TEXT
- embedding: VECTOR
  TYPE = FLOAT32
  DIM = 512
  INDEX_METHOD = HNSW
  DISTANCE_METRIC = COSINE
```

Listing 2. Image Index Schema Definition



This embedding is then compared to stored vectors in the text index using cosine similarity leveraging Redisearch's vector capabilities. The top-k most relevant articles (typically $k = 3$) are retrieved based on similarity scores. If a user uploads an image or screenshot, the system processes it using the ViT-B/32-quickgelu model via OpenCLIP to obtain a normalized 512-dimensional embedding. This embedding is queried against the image index, and top-matching results are returned using the same ANN-based retrieval mechanism. For both query types, the system supports Retrieval-Augmented Generation (RAG) by combining the retrieved context with the original user query. The retrieved content is formatted into a structured prompt that is passed to a generative LLM (e.g., GPT-4). Predefined instructions ensure that the model produces accurate, grounded answers based strictly on the retrieved information.

The generated response includes:

- A concise and semantically accurate answer to the user's query
- Inline or endnote-style references linking to the relevant source articles
- Confidence and traceability through alignment with original data

This RAG pipeline significantly reduces hallucinations by anchoring the generative process to high-quality, contextually relevant embeddings, while also enhancing transparency by linking answers to verifiable sources [8]. At inference time, the system seamlessly integrates multimodal retrieval and generation by embedding the user query, retrieving the most relevant content from Redis, and synthesizing an accurate, reference-backed response using a large language model [9].

3. RESULTS AND DISCUSSION

The implementation of a comprehensive, multimodal knowledge management system demonstrates substantial benefits for enterprise environments. Digital knowledge management systems provide rich affordances for organizational knowledge work, such as improved organizational memory and information sharing [10]. These systems enable better collaboration, productivity gains, and enhanced safety in workplace environments [11]. Furthermore, the integration of multimodal archive resources, including text, images, audio, and video, has become increasingly important for effective knowledge management in the era of big data [12]. By leveraging Large Language Models (LLMs)

Table 1. Benchmark results

Supports Unstructured Data	Traditional Search (Keyword-Based)	ML-Based Classifiers	Semantic Search (Text-Only)	Proposed Multimodal RAG Framework
Supports Unstructured Data	Limited	Moderate	High	High
Handles Image Queries	No	No	No	Yes
Contextual Understanding	Low	Medium	High	Very High
Summarization & QA Capabilities	No	No	Limited	Yes (via LLM)
Real-Time Semantic Retrieval	No	Slow	Yes	Yes
Scalability	Medium	Medium	High	High
User Satisfaction	Low	Medium	High	Very High
Ease of Integration	High	Medium	Medium	High

Table 2. Qualitative Assessments of Indexing Solutions for Semantic and Vector Search

Feature / Capability	Elasticsearch + kNN	FAISS	Pinecone	Redis + Redisearch
In-Memory Performance	Moderate	High	High	Very High
Real-Time Updates	Limited	No	Yes	Yes
Full-Text Search Integration	No	No	Limited	Yes
Vector Search Support	Plugin-Based	Native	Native	Native (via Module)
Horizontal Scalability	Medium	Requires Custom	Built-In	Built-In
Ease of Integration	Moderate	Complex	Easy	Easy
Deployment Flexibility	Self-Hosted Only	Self-Hosted Only	Cloud Only	Cloud or Self-Hosted
Multi-Modal Support	Requires Custom Code	Requires Custom	Moderate	Yes



and Vision Transformers (ViT), the system successfully converts unstructured data into a structured and semantically searchable format. This transformation is key to enabling more accurate information retrieval and enhancing decision-making processes across departments.

To evaluate the effectiveness of the proposed framework, a comparison was conducted against common retrieval methods across several key dimensions relevant to enterprise knowledge management.

As shown in Table 1, the proposed multimodal RAG framework consistently outperforms keyword-based search and single-modality systems across multiple dimensions, including contextual understanding, real-time retrieval, user satisfaction, and the ability to process both text and image inputs. The integration of question generation and article summarization using LLMs significantly enhances the accessibility of information and the speed with which users can make informed decisions [7].

In addition to evaluating retrieval methods, it is also important to consider the underlying infrastructure supporting semantic search. Table 2 compares Redis with other commonly used indexing and vector storage solutions in terms of speed, scalability, flexibility, and ease of integration.

As shown in Table 2, Redis with RediSearch provides several advantages over alternative indexing and vector database solutions, including in-memory speed, support for real-time updates, full-text search capabilities, and seamless integration of vector and scalar data. These characteristics are essential for enterprise applications that require low-latency access to large and evolving knowledge bases [5].

The combination of these two elements, a robust multimodal retrieval pipeline and a high-performance indexing backend, positions this system as a competitive and future-proof solution for enterprise knowledge management. It effectively bridges the gap between unstructured data and actionable insight, supporting scalable, cost-efficient operations while improving user satisfaction and organizational intelligence.

From a practical standpoint, the system has demonstrated strong scalability, with the ability to handle increasing volumes of data and user traffic without compromising performance. This makes it well-suited for use in large and growing enterprise environments. Moreover, the system's ability to reduce time spent manually searching for documents results in measurable cost savings, enhances user satisfaction, and reduces the risk of knowledge loss. Compared to traditional or single-

modality retrieval systems, the proposed multimodal RAG assistant provides a holistic and future-ready solution to enterprise knowledge management challenges.

4. CONCLUSION AND FUTURE IMPROVEMENTS

This work presented a robust and scalable multimodal knowledge management framework that leverages advanced machine learning models, specifically Large Language Models (LLMs) and Vision Transformers (ViT), in conjunction with a high-performance vector database to transform unstructured data into an accessible, searchable, and actionable resource. The system supports both text and image-based queries, enabling diverse information retrieval scenarios within enterprise environments. By integrating semantic embeddings, question generation, and article summarization, the framework significantly enhances decision-making, improves knowledge accessibility, and reduces the time spent searching for critical information. Furthermore, the modular design allows for horizontal scalability, cost efficiency, and seamless adaptation to evolving enterprise needs.

The proposed Multimodal Retrieval-Augmented Generation (RAG) Knowledge Assistant offers significant value for enterprise environments by transforming unstructured data into accessible, searchable knowledge [13]. It enhances decision-making by delivering accurate, context-rich information quickly and supports both text and image queries for comprehensive retrieval. Integration with large language models enables summarization and question-answering, reducing information overload and boosting productivity. By recovering lost or overlooked content, improving user satisfaction, and scaling with organizational growth, the system drives cost efficiency and provides a strategic advantage in managing enterprise knowledge.

Future enhancements aim to further increase the system's versatility and effectiveness by expanding support for additional file types such as PDFs and spreadsheets, enabling more comprehensive content search. Image retrieval will be improved through segmentation-based models that identify and focus on key visual regions, enhancing result relevance. The system will also support interactive file delivery within the chat interface, allowing users to download referenced content directly. Additionally, dynamic directory structuring will organize articles into semantic categories for better navigation. Ongoing model evaluation and exploring alternatives like BERT and DeepSearch, will optimize performance across various retrieval tasks and domains.



In summary, the continued development of this multimodal knowledge management framework will not only enhance its technical capabilities but also solidify its role as a strategic enabler of enterprise intelligence. By bridging unstructured data with accessible, context-rich insights, the system lays a strong foundation for future innovations in information retrieval, organizational learning, and scalable decision support.

REFERENCES

- [1] I. Aviv, I. Hadar, and M. Levy, "Knowledge Management Infrastructure Framework for Enhancing Knowledge-Intensive Business Processes," *Sustainability*, vol. 13, no. 20, p. 11387, Oct. 2021, doi: 10.3390/su132011387.
- [2] D. Li and Z. Zhang, "MetaQA: Enhancing human-centered data search using Generative Pre-trained Transformer (GPT) language model and artificial intelligence.," *PLoS One*, vol. 18, no. 11, p. e0293034, Nov. 2023, doi: 10.1371/journal.pone.0293034.
- [3] S. Sarto, L. Baraldi, A. Nicolosi, R. Cucchiara, and M. Cornia, "Towards Retrieval-Augmented Architectures for Image Captioning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 8, pp. 1–22, Jun. 2024, doi: 10.1145/3663667.
- [4] O. Frieder, C. I. Muntean, I. Mele, R. Perego, F. M. Nardini, and N. Tonello, "Caching Historical Embeddings in Conversational Search," *ACM Transactions on the Web*, vol. 18, no. 4, pp. 1–19, Oct. 2024, doi: 10.1145/3578519.
- [5] D. S. Asudani, P. Singh, and N. K. Nagwani, "Impact of word embedding models on text analytics in deep learning environment: a review.," *Artif Intell Rev*, vol. 56, no. 9, pp. 10345–10425, Feb. 2023, doi: 10.1007/s10462-023-10419-1.
- [6] Y. Yang, Y. Zhou, W. Shi, H. Fu, and S. Gao, "Intrusion detection: A model based on the improved vision transformer," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 9, Apr. 2022, doi: 10.1002/ett.4522.
- [7] X. Αποστολακη, "Extensive performance evaluation of popular relational and non-relational data stores for full-text search," 2025.
- [8] C. Yao and S. Fujita, "Adaptive Control of Retrieval-Augmented Generation for Large Language Models Through Reflective Tags," *Electronics (Basel)*, vol. 13, no. 23, p. 4643, Nov. 2024, doi: 10.3390/electronics13234643.
- [9] K. Rangan and Y. Yin, "A fine-tuning enhanced RAG system with quantized influence measure as AI judge," *Sci Rep*, vol. 14, no. 1, Nov. 2024, doi: 10.1038/s41598-024-79110-x.
- [10] H. Safadi, "Balancing Affordances and Constraints: Designing Enterprise Social Media for Organizational Knowledge Work," *MIS Quarterly*, vol. 48, no. 1, pp. 347–374, Mar. 2024, doi: 10.25300/misq/2023/16499.
- [11] J. Schuir and F. Teuteberg, "Understanding augmented reality adoption trade-offs in production environments from the perspective of future employees: A choice-based conjoint study," *Information Systems and e-Business Management*, vol. 19, no. 3, pp. 1039–1085, May 2021, doi: 10.1007/s10257-021-00529-0.
- [12] Y. Zhou, X. Wang, R. Zhao, Z. Zhang, and Q. Sheng, "Multimodal archive resources organization based on deep learning: a prospective framework," *Aslib Journal of Information Management*, Jan. 2024, doi: 10.1108/ajim-07-2023-0239.
- [13] S. Siriwardhana, S. Nanayakkara, R. Weerasekera, R. Rana, T. Kaluarachchi, and E. Wen, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Trans Assoc Comput Linguist*, vol. 11, pp. 1–17, Jan. 2023, doi: 10.1162/tacl_a_00530.



OPTIMIZATION OF KUBERNETES: RESOURCE ALLOCATION AND DYNAMIC SCALING

Džemil Sejdija,
[0009-0009-2042-0559]

Aldina Avdić*
[0000-0003-4312-3839]

State University of Novi Pazar,
Novi Pazar, Serbia

Abstract:

Kubernetes, a leading container orchestration platform, has become essential for managing modern cloud-native applications due to its scalability, automation, and resource optimization capabilities. This research focuses on Kubernetes' architecture, resource allocation strategies, and autoscaling mechanisms, highlighting key features such as the Horizontal Pod Autoscaler (HPA) and Vertical Pod Autoscaler (VPA). Through an analysis of experimental data and related works, the research underscores the importance of advanced scheduling algorithms, efficient monitoring tools like Prometheus and Grafana, and proactive resource management in improving overall operational efficiency. The findings demonstrate that combining Kubernetes-native features with customized enhancements can significantly reduce latency, resource contention, and operational costs, making Kubernetes a powerful tool for distributed application management.

Keywords:

Kubernetes, Resource allocation strategies, Autoscaling mechanisms, Horizontal Pod Autoscaler (HPA), Vertical Pod Autoscaler (VPA).

INTRODUCTION

Kubernetes (commonly abbreviated as K8s) is an open-source container orchestration platform designed to automate the deployment, scaling, and management of containerized applications. Developed originally by Google and later adopted by the Cloud Native Computing Foundation (CNCF), Kubernetes has become a cornerstone of cloud-native architectures. Its robust ecosystem enables developers to focus on building and delivering applications, while Kubernetes handles complex operational tasks such as load balancing, resource allocation, and service discovery.

Kubernetes achieves this through its distributed architecture, where multiple nodes operate as a unified cluster, ensuring high availability and fault tolerance. The control plane manages the cluster's state by monitoring workloads and scheduling tasks across worker nodes, which run the containerized applications. Pods, the smallest deployable units in Kubernetes, encapsulate one or more containers and share network and storage resources within the same namespace. This architecture supports horizontal scaling, where additional pods can be deployed dynamically based on demand, and vertical scaling, where resource allocations for existing pods can be adjusted.

Correspondence:

Aldina Avdić

e-mail:

apljaskovic@np.ac.rs





With the rapid growth of cloud-native computing, Kubernetes has become indispensable for organizations seeking to deploy resilient, scalable applications across hybrid, multi-cloud, or edge environments. However, as the scale and complexity of Kubernetes clusters grow, challenges such as resource underutilization, high costs, and performance bottlenecks necessitate adopting advanced optimization techniques. This research delves into Kubernetes' core components and resource management capabilities, focusing on strategies for optimizing autoscaling, dynamic resource allocation, and cost efficiency to enhance system performance and reliability in cloud-based environments. The methodology used in this work is experimental. The paper consists of an overview of similar works, an overview of Kubernetes architecture, an overview of Kubernetes resources and optimization methods, actual experiments for different techniques, and a discussion of results.

2. LITERATURE OVERVIEW

Experimental data from industry examples demonstrate the benefits of optimized Kubernetes configurations.

In the paper [1] key topics are Kubernetes autoscaling mechanisms (HPA, VPA and CA) and performance evaluation of HPA using Prometheus. Experimental insights provided by the paper and practical lessons show how to enhance the efficiency of resource management in the Kubernetes environment.

In the paper [2], key topics are performance bottlenecks (inefficient autoscaling) and experimental results that show performance improvements. The focus on reducing latency and enhancing scheduling strategies achieves efficient dynamic resource allocation.

In the paper [3], the author uses techniques for optimizing resource allocation but also incorporates predictive analytics to anticipate workload demands. The relevance of this work is the combination of Kubernetes-native features with third-party tools. In combination with other works, they give a lot of information regarding: Resource Optimization, Autoscaling Insights, Performance Improvement and Tool integration (Prometheus).

3. ARCHITECTURE OF KUBERNETES

The architecture of Kubernetes (Figure 1) follows a master-worker distributed model, designed to efficiently manage containerized workloads across multiple nodes. This architecture ensures scalability, reliability, and fault tolerance. The primary components are divided into control plane components (master node) and data plane components (worker nodes), with each serving a distinct purpose to maintain the desired state of the cluster [4] [5].

3.1. THE CONTROL PLANE (MASTER NODE)

The control plane is responsible for managing the overall state of the cluster, ensuring that the desired configuration is maintained. The key components are:

- **API Server:** The communication hub of Kubernetes, handling requests and updates from users and internal components.
- **Controller Manager:** Ensures the cluster stays in the desired state by managing various controllers (e.g., replication and node health).
- **Scheduler:** Assigns workloads (pods) to suitable worker nodes based on available resources and constraints.
- **etcd:** A distributed storage system that holds cluster configuration and state data, ensuring consistency and fault tolerance.

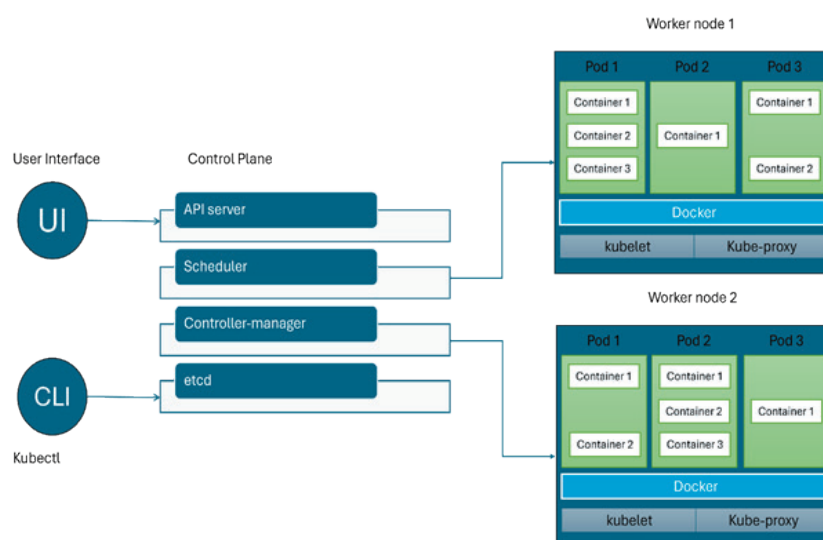


Figure 1. An Architecture of Kubernetes



3.2. THE DATA PLANE (WORKER NODES)

Worker nodes run the applications and provide the necessary computing resources. The key components are:

- kubelet: An agent on each node that ensures the assigned containers are running and healthy.
- kube-proxy: Manages network communication within the cluster and to external clients.
- Container Runtime: Responsible for running containerized applications (e.g., Docker, Containerd).

3.3. CORE KUBERNETES OBJECTS

Kubernetes uses objects like Pods, ReplicaSets, Deployments, Services, and Ingress to manage containerized workloads, ensuring scalability, availability, and network accessibility within and outside the cluster. Additionally, features like Namespaces enable logical isolation for multi-tenancy, Horizontal Pod Autoscaler (HPA) scales pods based on resource usage, and Vertical Pod Autoscaler (VPA) optimizes resource allocation dynamically.

Kubernetes facilitates communication within the cluster through service discovery and internal DNS. When a pod is created, it is assigned a unique IP address within the cluster. Services enable other pods or external clients to access these pods via stable DNS names, avoiding direct dependency on pod IPs that may change over time [6] [7].

4. KUBERNETES RESOURCES AND OPTIMIZATION METHODS

Efficient resource management in Kubernetes is crucial for achieving high performance, cost efficiency, and system reliability. Kubernetes provides various mechanisms to allocate and manage CPU, memory, and storage resources across containerized applications. It offers both static and dynamic resource allocation, allowing workloads to scale based on demand.

When it comes to CPU management, Kubernetes ensures that containers receive a guaranteed minimum amount of CPU, while also setting an upper limit on how much they can consume. For example, if a container requests a fraction of a CPU core, it is guaranteed that amount, but it cannot exceed a specified limit.

Memory management works in a similar way, where a container is allocated a minimum amount of memory to ensure stable performance. However, if it exceeds the defined limit, it may be terminated to prevent excessive resource consumption.

In terms of storage, Kubernetes supports different types of storage solutions. Persistent storage can be either pre-provisioned or dynamically created based on application needs, ensuring data remains available even if a pod is restarted. Applications can request specific storage capacity using claims, while ephemeral storage provides temporary space that exists only for the duration of a pod's lifecycle.

By leveraging these resource management features, Kubernetes ensures that applications run efficiently, using resources optimally while maintaining system stability [8] [9].

Kubernetes optimizes resource use through autoscaling, quotas, and monitoring. Autoscaling adjusts resources based on demand—scaling pods horizontally to handle increased load, vertically adjusting resource limits (sometimes requiring restarts), and scaling clusters by adding or removing nodes to control costs.

Resource management is enforced with quotas and limits. Quotas cap total CPU, memory, and storage in a namespace, while limit ranges set default resource allocations to prevent over or under-provisioning.

Workload placement is optimized using affinity rules to assign pods to specific nodes and taints/tolerations to isolate resource-heavy workloads.

Monitoring tools like Prometheus, Grafana, and kube-state-metrics provide real-time insights into resource usage, helping ensure efficient performance and cost management [10] [11] [12].

5. EXPERIMENTS

5.1. EXPERIMENT 1: HPA TESTING FOR CPU UTILIZATION

The objective is to observe HPA behavior in response to increasing CPU utilization and pod scaling. The cluster should have three worker nodes (4 CPUs, 8 GB RAM each). Prometheus should be used for monitoring. The methodology is as follows: deploy nginx (CPU 250m request, 500m limit, memory 256Mi request, 512Mi limit). Configure HPA to scale based on CPU and run a load test with 2000req/sec using "hey". The results are in Table 1.



5.2. EXPERIMENT 2: VPA RESOURCE ADJUSTMENT

The objective is to analyse how VPA adjusts CPU and memory allocations for pods based on usage patterns. The cluster should have three worker nodes (4 CPUs and 16 GB RAM). The methodology is: to deploy a Linux image and make it do a batch processing task every 5 min (it can only generate and discard 128-256 MB of data to simulate CPU and memory load). Set initial CPU request to 100m and memory to 128Mi. Enable VPA and monitor how VPA adjusts resources. The results are in Table 2.

5.3. EXPERIMENT 3: RESOURCE QUOTAS AND COST CONTROL

The objective is to enforce resource quotas and monitor cost savings by restricting excessive resource usage across namespaces. The cluster should have four nodes (4 CPUs and 8 GB RAM) and two namespaces (team-a and team-b). Apply resource quotas to both namespaces. Deploy applications that attempt to request higher resources. Motor allocation is done using Grafana. The results are in Table 3.

6. DISCUSSION

Experiment 1 showcased how the Horizontal Pod Autoscaler (HPA) dynamically adjusted the pod count based on CPU usage. When usage exceeded 50%, the HPA increased the number of pods to handle the load, preventing performance issues. As the load decreased, the HPA scaled down the pods, optimizing resource usage and reducing costs. This demonstrated the HPA's effectiveness in maintaining system stability and ensuring responsiveness during peaks while conserving resources during idle times. Overall, the HPA proved to be a valuable tool for balancing performance and cost efficiency.

Experiment 2 demonstrated how the Vertical Pod Autoscaler (VPA) adjusted CPU and memory based on workload demands. As resource usage increased, VPA allocated more resources to ensure smooth performance. Conversely, when the demand dropped, it scaled down resource allocation to avoid over-provisioning. By dynamically adjusting resources, VPA optimized allocation, preventing inefficiencies, and ensuring the system operated cost-effectively.

Experiment 3 demonstrated how resource quotas effectively managed resource usage by enforcing limits. When team-a exceeded their allocated quota, their excessive requests were throttled, ensuring that other teams had fair access to resources.

Table 1. HPA Results

Time	CPU% Utilization	Number of pods
0	20	2
5	55	4
10	70	6
15	80	8
20	48	4

Table 2. VPA Results

Time	CPU Request(m)	Memory Request
0	100	128
5	200	256
10	400	512
15	250	384
20	150	256

Table 3. Resource quota results

Namespace	Requested CPU	Allocated CPU	Status
Team-a	3	2	Limited
Team-b	1.5	1.5	Allowed



By implementing quotas, resource allocation remained balanced and prevented any team from monopolizing resources, promoting fairness across the system.

7. CONCLUSION

Effective resource management in Kubernetes is further supported by proactive monitoring through tools such as Prometheus and Grafana. These monitoring solutions enable real-time visibility into resource consumption, providing valuable insights that allow for timely adjustments and informed decision-making. By continuously tracking key performance metrics, organizations can identify potential bottlenecks before they impact operations, ensuring that resources are utilized efficiently, and applications remain stable under varying workloads. The integration of monitoring with autoscaling capabilities contributes to a more resilient and responsive infrastructure, ultimately improving the reliability and efficiency of cloud-native applications. The combination of HPA and VPA offers a holistic approach to scalability and resource management within Kubernetes environments. While HPA provides the ability to scale horizontally by increasing or decreasing the number of running pods based on load, VPA fine-tunes resource allocations within individual pods to ensure efficient utilization. This synergistic approach enhances the responsiveness of applications to fluctuating workloads and ensures that resources are allocated precisely where they are needed, reducing waste and improving overall performance. Together, these autoscaling mechanisms provide a comprehensive solution to the challenges of managing cloud-native applications in dynamic environments.

REFERENCES

- [1] T.-T. Nguyen, A. M. Rahman, Y. H. Tran, Q. M. Tran, and C. H. Choi, "Horizontal pod autoscaling in Kubernetes for elastic container orchestration," *Sensors*, vol. 20, no. 16, p. 4621, 2020. doi: 10.3390/s20164621.
- [2] S. K. Mondal, Z. Zheng, and Y. Cheng, "On the optimization of Kubernetes toward the enhancement of cloud computing," *Mathematics*, vol. 12, no. 16, p. 2476, 2024. doi: 10.3390/math12162476.
- [3] A. Mustyala, "Dynamic resource allocation in Kubernetes: Optimizing cost and performance," *EPH - International Journal of Science and Engineering*, vol. 7, no. 3, pp. 59–71, 2021.
- [4] C. C. Chang, S. R. Yang, E. H. Yeh, P. Lin, and J. Y. Jeng, "A Kubernetes-based monitoring platform for dynamic cloud resource provisioning," in *Proc. IEEE Global Communications Conference (GLOBE-COM)*, Singapore, 2017, pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/document/8254046>
- [5] G. Rattihalli, M. Govindaraju, H. Lu, and D. Tiwari, "Exploring potential for non-disruptive vertical auto scaling and resource estimation in Kubernetes," in *Proc. IEEE Int. Conf. on Cloud Computing (CLOUD)*, Milan, Italy, 2019, pp. 33–40. [Online]. Available: <https://ieeexplore.ieee.org/document/8814504>
- [6] M. Song, C. Zhang, and H. E, "An auto scaling system for API gateway based on Kubernetes," in *Proc. 2018 IEEE 9th Int. Conf. on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2018, pp. 109–112. doi: 10.1109/ICSESS.2018.8663784.
- [7] Y. Jin-Gang, Z. Ya-Rong, Y. Bo, and L. Shu, "Research and application of auto-scaling unified communication server based on Docker," in *Proc. 2017 10th Int. Conf. on Intelligent Computation Technology and Automation (ICICTA)*, Changsha, China, 2017, pp. 152–156. [Online]. Available: <https://ieeexplore.ieee.org/document/8089924>
- [8] P. Townend, A. Basu, M. Eisa, and J. Kołodziej, "Improving data center efficiency through holistic scheduling in Kubernetes," in *Proc. 2019 IEEE Int. Conf. on Service-Oriented System Engineering (SOSE)*, Newark, CA, USA, 2019, pp. 156–166. [Online]. Available: <https://ieeexplore.ieee.org/document/8705815>
- [9] D. Balla, C. Simon, and M. Maliosz, "Adaptive scaling of Kubernetes pods," in *Proc. IEEE/IFIP Network Operations and Management Symposium (NOMS)*, Budapest, Hungary, 2020, pp. 8–12. [Online]. Available: <https://ieeexplore.ieee.org/document/9110428>
- [10] E. Casalicchio and V. Perciballi, "Auto-scaling of containers: The impact of relative and absolute metrics," in *Proc. 2017 IEEE 2nd Int. Workshops on Foundations and Applications of Self Systems (FASW)*, Tucson, AZ, USA, 2017, pp. 207–214. [Online]. Available: <https://ieeexplore.ieee.org/document/8064125>
- [11] J. Santos, T. Wauters, B. Volckaert, and F. D. Turck, "Resource provisioning in fog computing: From theory to practice," *Sensors*, vol. 19, no. 10, p. 2238, 2019. doi: 10.3390/s19102238.
- [12] W. S. Zheng and L. H. Yen, "Auto-scaling in Kubernetes-based Fog Computing platform," in *Proceedings of the International Computer Symposium*, Singapore: Springer, 2018, pp. 338–345. doi: 10.1007/978-981-13-9190-3_35.



STUDENT SESSION

GAME DEVELOPMENT – HOW TO CREATE AN IMMERSIVE STORYTELLING EXPERIENCE

Nikola Peruničić*,
[0009-0004-0331-9938]

Mlađan Jovanović
[0000-0003-2355-9424]

Singidunum University,
Belgrade, Serbia

Abstract:

Video games have seen steady growth in recent years. The market value of the industry was valued at more than 271 billion dollars and boasted an impressive 3.32 billion active player base. When talking about the gaming industry, there are two main types of genres: AAA (triple-A) games and Indie games. AAA is associated with the biggest corporations and budgets. Companies such as Ubisoft, Electronic Arts, Nintendo, and Valve fit the classification. Their games have huge budgets and are well-known in the industry. On the other end of the spectrum, indie games can be found, which are created by small companies or individuals. In recent years, the increased accessibility of game engines has allowed the indie genre to experience significant growth. This paper is focused on creating an engaging experience in the video game medium.

Keywords:

Video Games, Game Engines, Heaven Valley, Storytelling.

INTRODUCTION

Video games carry the possibility to immerse a player in ways many other mediums cannot. The potential in certain combinations of world-building, level design, and gameplay offers a unique experience that otherwise could not be recreated. Most games we see today have their roots in indie development. Indie games are a derivative of independent video games or games made by individuals or a small team. A lot of popular games we see today started as a hobby led by enthusiasts. For example, Minecraft, which is now owned by Microsoft and is one of the most popular games out there, started as an indie game. Thirty years ago, the people responsible for electronic entertainment were not yet game professionals; they were dedicated hobbyists [1]. A simple hobby has led to a multimillion-dollar industry with an ever-growing consumer base.

One of the more interesting ways to create a story inside of a game is to utilize dreams. Dreams are a highly subjective phenomenon, subject to a plethora of interpretations, coming from both formal schools of psychology and popular, folkloric interpretations.

Correspondence:

Nikola Peruničić

e-mail:

nikola.perunicic.mejl@gmail.com





This creates opportunities in level design that would not make sense otherwise. Along with level design, items that occur in dreams can be used gameplay-wise to create an interesting narrative and fun mechanics. Sigmund Freud theorized that all material making up the content of dreams is in some way derived from experience [2]. Items that occur in said dreams could relate to something that happens in real life. For example, a jar full of teeth could symbolize the loss of a loved one or something you hold precious. This can be used to create a narrative about the main character based on what items are put inside the game. *Heaven Valley* uses this and furthers it by creating a dream journal that keeps track of all picked-up items and their descriptions. Along with this, some items give the player a power-up that will help him in his journey. This adds an RPG (*role-playing game*) element to *Heaven Valley*.

The concept of an RPG consists of a game in which players assume the roles of characters in a fictional setting. The original term comes from the Dungeons and Dragons tabletop *role-playing game* (*TTRPG* or *TRPG*). The way this translates into a game is by having a role-playing aspect (characters you meet and play), while the focus is on the storytelling element and character progression system.

2. WHAT IS A GAME?

A Game is a system of rules in which agents compete by making ambiguous decisions [3]. The system of rules means a guideline on how the game is supposed to be played and how the agents refer to the players. There is another definition that closely correlates with the first one. A game is a problem-solving activity approached with a playful attitude [4].

2.1. DESIGNING WITH UNITY

Game engines are software platforms that allow developers to create games efficiently. Popular examples include Unreal Engine, Unity, and Godot. This paper will focus on the Unity game engine. Unity supports a variety of platforms, like desktop, mobile, console, augmented reality, and virtual reality. Its main source code is C# (C sharp). Unity is a game development ecosystem: a powerful rendering engine fully integrated with a complete set of intuitive tools and rapid workflows to create interactive 3D content; easy multiplatform publishing; thousands of quality, ready-made assets in the Asset Store; and a knowledge-sharing Community [5].

2.2. STORY IMPLEMENTATION

All complete stories have three different phases [6].

- Beginning: Setup;
- Middle: Confrontation; and
- End: Resolution.

The beginning serves as an introductory point to the game. The way it is established is important to keep the player hooked to the game. If conveyed properly, the overall tone and atmosphere can be felt from the very beginning to the end. The goal of the first phase is to immerse the player and guide him towards the end of the game.

The middle phase is where most of the conflict and story will occur. Its goal is to pay off what has happened in the beginning and to set up the finale. Given that is the biggest part of the game, it is very easy for a developer to lose track and progression, which will lead to an unsatisfactory gameplay experience.

The end serves as the peak and the resolution of the game. It is here where the player can encounter the final boss or the last piece of the puzzle of the story. It can be accompanied by multiple endings based on what the player did throughout the game. The developer can make the first two phases perfect, but if he does not do the last phase correctly, it can lead the player to have a bitter experience after completing the game.

2.3. COMBINING STORY WITH GAMEPLAY MECHANICS

When developing a game story, you will find there are three different types of people in your audience.

- Players that are partially invested in the story;
- Players who are fully invested in the story; and
- Players that are not interested in the story.

The best rule of thumb is to always make the story in service of the gameplay, not the other way around [7].

The way *Heaven Valley* goes about this is by creating an inventory system labeled as a "dream log". The purpose of the dream log is to give the player pieces of the story by collecting items on various levels. All picked-up items will have a description that will sometimes hint at what the story may be about. Some items have a negative impact and affect the outcome of the game. Some items are hidden, and they serve as a mechanic to reward the player for going out of his way to search for such items. Along with this, the character will comment on anything unusual they notice. Earlier games like *Penumbra* used to have the main character give comments on almost



any item that could be interacted with. *Half-life* had *non-playable characters* (NPCs) that would comment on what the player was doing during the gameplay or if the NPC interacted with. Such small details can immerse the player further and make the world feel more alive.

2.4. AUDIO DESIGN

Audio Design refers to the creation of sound elements inside the game. For instance, if the developer wants to have his characters speak, it would be good practice to hire professional voice actors. A good voice actor can elevate and better present the personality of any character he voices. Creating a soundtrack for each level or sequence will set the tone and atmosphere of the game. Audio soundtracks help establish and embellish the narrative setting [8]. Ambient noises like wind, water, and animals contribute to the atmosphere, making the world feel more alive. Audio design is essential for creating an engaging and immersive experience.

2.5. LEVEL DESIGN

The primary function of a game environment is to guide the player's movement [9].

Level Design is the process of creating the environment of the game. It is a crucial part of game development because it is here where everything about the game will be shown. The visual design needs to correlate with the theme of the game. It is important to make sure that the theme is reflected throughout several areas of the level, or it will not have the impact it should have [10]. Lighting the environment can add depth to any level. It can hide objects in the background to blend in more with the level, or it can increase the effects of details. Lights can be employed to indicate a goal, a path towards an exit, or a point of interest [11].

Visual design is usually accompanied by audio design. Likewise, audio design needs to follow a certain theme to further the visual impact presented before the player.



Figure 1. Example of interactive storytelling in *Penumbra: Black Plague*



Figure 2. Example of interactive storytelling in *Heaven Valley*



3. HEAVEN VALLEY LEVEL DESIGN AND HOW THE THREE PHASES ARE INCORPORATED

The first level in *Heaven Valley* is called the “Forest” level. It serves as an introductory level to get the player used to the controls and initial mechanics. The player starts at a bus station. The main character comments on how her mother asked her to go get the groceries across the road. This gives the player the first task he needs to fulfill to progress. After reaching the store, the player discovers that it has been abandoned for a long time, and the main character comments on this. This creates intrigue in the environment and will prompt the player to explore his surroundings to help him understand the mystery behind what is happening. The world will start making less sense, and eventually, it will be obvious that the main character is no longer in the real world but in a dream.

The conclusion of the initial phase comes after the player finishes the first level and enters the second level called “Throne room”. In this level, the player will meet an NPC that will give him the end goal of collecting 3 different items and the end destination where the player will be “judged”. Upon exiting the Throne room, the player will enter the “Hills” level and will thus finish the initial phase and go towards the middle phase.

The middle phase utilizes dreams in level design to a greater extent. After exploring the “Hills” level, you will find that there is a locked brutalist building. After finding the key and entering the building, the player will enter the next level called “Brutalist Verticality”.

The world that had some sense before now loses all of it. The player will find himself surrounded by concrete with rusted rebar sticking out of it. The surrounding area is covered with fog, and the atmosphere is claustrophobic. This level features a platforming section that is made easier if the player has collected the “bike” item that gives him a speed boost. In the distance,



Figure 3. Brutalist Verticality Level



Figure 4. Good ending level



a huge building can be seen. This is the end destination the NPC at the “Throne room” level was referencing. After completing the platforming section and reaching the building. The player will enter the next level called “Tower”. This is where the player will be “judged”. Depending on the player's actions and what items he collected, he will receive one of the two endings.

The finale of *Heaven Valley* is concluded in the “Dream Library” level. If the player collects every item and completes every quest, he will get a good ending. Failure to do so will result in a bad ending.

4. DISCUSSION

The main inspiration for *Heaven Valley* is a game called *Yume Nikki*. *Yume Nikki* is a two-dimensional (2D) RPG game made in the RPG Maker engine. It is one of the most well-known indie games that came out of Japan. It is an exploration game in which the player can explore a set of bizarre and dreamlike worlds. In *Heaven Valley*, you can see a full 3D environment [12]. This allows for more implementation of mechanics, level details such as fog and rain, and real-time lighting for realistic shadows. Compared to *Yume Nikki*, such features are not possible because of its 2D nature.

4.1. STORYTELLING IN COMPARATIVE GAMES

The way *Thief: The Dark Project* does its storytelling is through intro cutscenes, mission briefings, and environmental storytelling. You play as Garrett, and your first mission is stealing a priceless trinket from Lord Bafford's Manor. Along with the trinket, you are tasked with stealing as many valuable items as possible. The realistic way Manor is designed adds immersion to the game. You have your dining room, kitchen, sleeping quarters, and guest rooms. The upper floors are filled with luxurious pools, lavish bedrooms, and the lords' quarters. Lower areas do not contain a lot of valuable items. You can find an occasional coin pouch and silver cups, but to progress, the player will need to go to the upper level where the nobles reside. This adds an aspect of realism to the environmental storytelling. We do not expect the servants to possess much wealth. On the other hand, the nobles leave their coins and valuables wherever they please, making it easier to steal. This gives us a glimpse of how the societal hierarchy is structured inside of the game.

Heaven Valley is a dream-like world that has a semblance of reality. For instance, if the player enters the store in the “Forest” level, he can expect to find a bottle that has a blue label on it with the description saying it is cold to the touch. This could be a reference to anything the store was selling before it was abandoned. The resemblance to *Thief* becomes more apparent on later levels of the game, where Garrett is tasked with breaking into the mansion of the nobleman Constantine to steal a unique sword. The way the level is designed is supposed to lure the player into a false sense of security, thinking he is only stealing from a normal mansion. The moment the player starts exploring more, he will realize that the rooms make no sense. Most of them will be turned upside down, resemble miniature versions of existing rooms, or lead to completely unexpected rooms like a garden or a tomb. This unexpected transition happens in *Heaven Valley*, too. Most notably, when the player in the “Hills” level enters a brutalist building, and ends up inside a vast concrete landscape with huge buildings surrounding him, as shown in Figure 3.

Another inspiration for *Heaven Valley* comes from a game called *Fatum Betula*. The introductory sequence begins at the church level, where the player is guided to ascend the staircase and see a tree attached to a pool of water, with a vast hallway leading into darkness. Upon staring into the darkness, the player will notice an NPC approaching. The NPC explains that the tree represents life, and tainting the water beneath it will transform the world into a new one. The middle phase of *Fatum Betula* involves completing various quests given by other NPCs to collect as many different liquids as possible. The ending is achieved by tainting the water with the gathered liquids. The resemblance to *Heaven Valley* appears in various forms. The first similarity is with the main quest giver: in *Fatum Betula*, it is a monster-like NPC in the first level, while in *Heaven Valley*, it is the “Headless Statue” NPC in the second level. Another similarity lies in level design, where both games feature dream-like worlds and states. Inside *Fatum Betula*, the player can poison the sea to eliminate the vines guarding the deeper levels. Upon venturing deeper into a newly unlocked area, the player will find themselves in a new level resembling an American suburban area. Further exploration leads to a house reminiscent of the main character's childhood home. This sudden transition from the deep sea to a completely different area indicates that we, as players, are diving deeper into the psyche of the main character, further represented by a childhood memory encountered upon entering the house. Similarly, in *Heaven Valley*, the player can discover buildings and



items that evoke memories for the main character. The "Tower" level, where the player is to be judged, symbolizes that the main character is ready for a resolution, which, depending on the player's choices, will be fully resolved or not at all.

5. CONCLUSION

Game development is a very taxing process that can quickly stir a game away from its goal if taken lightly. Technological advancements in game engine software have allowed people to create games with relative ease. This has come with a downside in that a lot of indie games that are released are below players' expectations. The purpose of this research is to help future developers understand what makes games enjoyable and apply that to their work. *Heaven Valley* is one example of how to use thematic and environmental storytelling. It is included because playing the game should further drive the previously presented points, and it can be downloaded from GitHub [13]. Video games are a relatively new concept, and people are still adapting to the possibility of it being an art form. With continuous innovation and quality releases, this mentality is bound to change.

REFERENCES

- [1] T. Fullerton, *Game Design Workshop: A Play-Centric Approach to Creating Innovative Games*, Elsevier Inc., 2008.
- [2] S. Freud and J. Strachey, *The interpretation of dreams*, Basic Books, 2010.
- [3] K. Burgun, *Game Design Theory*, CRC Press, 2012.
- [4] J. Schell, *The Art of Game Design: A Book of Lenses*, Elsevier Inc., 2008.
- [5] R. Hill-Whittall, *The indie game developer handbook*, New York: Focal Press, 2017.
- [6] E. Skolnick, *Video game storytelling: what every developer needs to know about narrative techniques*, New York: WATSON-GUPTILL, 2014.
- [7] S. Rogers, *Level Up!*, John Wiley & Sons, 2010.
- [8] K. Salen and E. Zimmerman, *Rules of Play: Game Design Fundamentals*, Mit Press, 2004.
- [9] R. Kemppinen, "Designing Visuals for a Story Rich Game," B.S. thesis, Tampere University of Applied Sciences, 2019. [Online]. Available: <https://www.theseus.fi/handle/10024/185838>.
- [10] R. Kremers, *Level Design*, Taylor & Francis Group, 2009.
- [11] M. Salmond, *Video Game Level Design: How to Create Video Games with Emotion, Interaction and Engagement*, New York: Bloomsbury Academic, 2020.
- [12] Kozalog, "Heaven Valley walkthrough," YouTube, 28 February 2025. [Online]. Available: <https://www.youtube.com/watch?v=Kn1pWI2XrNQ>. [Accessed 7 April 2025].
- [13] N. Perunicic, "Github," 2 March 2025. [Online]. Available: <https://github.com/Kozalog/HeavenValley.git>. [Accessed 7 April 2025].



STUDENT SESSION

THE ROLE OF ARTIFICIAL INTELLIGENCE IN WASTEWATER TREATMENT

Milica Radojković*,
[0009-0001-8185-7211]

Milena Rikalović
[0000-0002-1809-5461]

Singidunum University,
Belgrade, Serbia

Abstract:

This paper analyses the role of artificial intelligence (AI) in wastewater treatment. It highlights how artificial neural networks (ANNs) can effectively process complex data, such as the correlation between biological and chemical parameters, which are often not linearly related. These models excel in handling large datasets, making them more effective than traditional methods for monitoring wastewater. Although AI is portrayed in media, it also serves as a valuable tool in (bio)technological processes, especially in complex systems that manage large datasets. The use of AI, particularly machine learning (ML), is growing in wastewater treatment plants (WWTP)s to optimise their operations. So far, AI implementation in this field has led to more effective data analysis results than those achieved through conventional statistical approaches.

Keywords:

Artificial Intelligence, Wastewater Treatment, Data Science, Machine Learning.

INTRODUCTION

In modern times, the importance of water as a natural resource and polluted water treatment is one of the most important topics in contemporary science and (bio)technology, considering environmental protection. The special emphasis is on industrial wastewater and the challenges faced by treatment plants due to extensive urbanisation and industrialization since the middle of the last century. This aligns with sustainable development goals and rational resource utilization, which is highlighted by increased water stress on a global level. [1]

Effective wastewater treatment is essential not only for maintaining a safe water supply but also for preserving ecosystems and safeguarding human health. Nevertheless, conventional treatment methods often prove inadequate in tackling the complex challenges presented by modern-day pollutants. [2]

Correspondence:

Milica Radojković

e-mail:

milica.radojkovic.24@singimail.rs





As a result, the reclamation and reuse of treated wastewater have become increasingly important strategies. This process is widely acknowledged as one of the most efficient solutions to the issue of water scarcity, playing a crucial role in the sustainable management of water resources by enabling reuse across various sectors. [3]

Therefore, it is imperative to improve and further develop wastewater treatment technologies by adopting modern methods. This need is driven by the escalating concern for water quality, which has become a significant global issue. These challenges are further intensified by current technological limitations in delivering comprehensive and reliable water quality. [3]

One of the modern approaches to this process is artificial intelligence (AI), which enables the analysis of large amounts of data, optimization of parameters, and prediction of results in real time. The concept of AI has been evolving since the 1950s, and in recent decades it has become increasingly popular thanks to advances in computer power and algorithms. [3]

AI is used in various areas, including planning, data mining, decision making, language processing, and even humanoid robots. Its application in wastewater treatment opens new possibilities for more efficient and sustainable management of this resource. [3]

The present study analyses and summarizes the present application of AI for wastewater treatment, as an important tool for this field in applied science. [3]

2. COMPONENTS OF ARTIFICIAL INTELLIGENCE

2.1. DATA

Data forms the core of artificial intelligence, which is capable of handling both quantitative and qualitative information. While AI can process a wide range of data types, such as text, images, and audio, the extent to which it can be applied to numerical or categorical data in wastewater treatment plants (WWTPs) is still under exploration. [4]

Quantitative data relate to precise numerical values and can be either continuous or discrete. In water treatment plants, most of the data is continuous, including flow rates, oxygen concentration, and energy use. [4]

Qualitative information includes details such as identification labels, the presence of faults, and the system's operational status (normal or abnormal). Typically, numerical data are presented in tabular format, where rows represent individual observations, and columns are referred to as attributes or variables that describe specific data characteristics. [4]

It can be stated that artificial intelligence (AI) operates as a model driven by information that is dependent on the quality of the information it processes. The greater the quantity of data available, the greater the accuracy of the AI design becomes – hence, extensive datasets are crucial for efficient training. When insufficient data is used, challenges such as overfitting the training data and poor performance on test data may arise. Issues such as the trained model's greater sensitivity to noise or high dimensionality often arise alongside these problems. [4]

In recent research, special emphasis has been placed on the data pre-processing and conditioning as a prerequisite for developing robust AI applications. This includes cleaning raw sensor inputs, normalising scales, and selecting relevant features that are crucial in dynamic systems like WWTPs. High-quality, structured, and representative data allow AI algorithms to effectively learn the underlying process behaviour and adjust predictions in real time, thereby improving operational control and resource optimisation. As such, the data are not only the fuel but also the framework for deploying successful AI models in complex treatment environments. [5]

2.2. ALGORITHMS

AI algorithms are commonly classified in various ways, including: supervised, semi-supervised, unsupervised, and reinforcement learning algorithm (depending on the learning process); linear and non-linear algorithms (based on the nature of the function they model); and approaches such as machine learning (ML) and deep learning (DL). [4]

In the context of WWTP algorithms, there are hundreds of AI models available, and the most suitable one for a specific issue is not always known beforehand. [6]

The most effective method for selecting an AI algorithm is through random sampling and the iterative trial-and-error method (TaE), which involves comparing multiple algorithms in parallel. [4]

Additionally, the selection of the algorithm is influenced by the nature of the data, the type of issue being tackled, and the intended results. [4]

Generally speaking, AI algorithms are diverse, with a wide range of models to choose from. It is not feasible to identify the optimal algorithm for a particular issue beforehand. [4] Moreover, the choice of algorithm depends on the nature of the data, the problem being addressed, and the desired outputs. [6]



Additionally, considerations such as computational efficiency, scalability, and interpretability of the model also play an important role in algorithm selection. For example, while complex models like deep neural networks may offer high accuracy, they often require greater computational resources and may be more difficult to interpret. Therefore, balancing model performance with practical implementation constraints is essential for achieving reliable and sustainable outcomes in real-world applications. [7]

Recent studies underscore that although deep learning architectures-particularly convolutional and recurrent neural networks-demonstrate considerable strength in pattern recognition and time-series forecasting, their black-box nature presents notable limitations in terms of explainability, which can be crucial in safety-critical settings such as WWTP (Tomar et al., 2019). [2] In comparison, more straightforward models like decision trees or support vector machines may yield slightly lower accuracy, yet they offer enhanced transparency and interpretability (Bukhari et al., 2021). [8] Therefore, the balance between model performance and interpretability should be assessed on a case-by-case basis, particularly when AI systems are employed to monitor chemical and biological parameters within WWTPs. [8]

3. EXAMPLES OF ARTIFICIAL INTELLIGENCE IN WASTEWATER TREATMENT

Artificial intelligence surged in popularity during the 1980s, and its applications have since expanded across various industries. However, utilizing AI to address urgent challenges in wastewater treatment remains complex, due to the need for specialized skills (e.g. computing systems), knowledge (e.g., data science and statistical methods), and resources (e.g., sufficient datasets). [4] Moreover, selecting an inappropriate problem can lead to the failure of AI implementation. Previous studies have utilised a variety of approaches, such as predicting one effluent quality indicator from another, incorporating data from different stages of the process, or combining influent quality metrics with operational factors, such as return sludge flow rate, sludge volume index, food-to-microorganism ratio, sludge retention time, and energy or chemical consumption. [4]

For example, Zhao et al. (2016) [6] created an artificial neural network (ANN) model aimed at forecasting the effluent levels of total phosphorus (TP), biological oxygen demand (BOD), chemical oxygen demand (COD), total suspended solids (TSS), and ammonium

nitrogen at a wastewater treatment facility located in China. The model was built using input data that included raw wastewater quality indicators-such as influent levels of total phosphorus (TP), biological oxygen demand (BOD), chemical oxygen demand (COD), total suspended solids (TSS), ammonium nitrogen ($\text{NH}_4\text{-N}$), and pH together with data on energy and chemical consumption such as electricity, coagulants, and flocculants. [9]

Russell & Norving (2016) [6] employed artificial intelligence, specifically a soft sensor built on neural network principles, designed to estimate effluent parameters (including COD, TN, TSS, O_2 , NO_3 , NH_4 , and alkalinity) and represent the non-linear dynamics of the wastewater treatment process. [4] This tool was also utilised for the optimal control of aeration, pumping, and disposal costs, while ensuring compliance with effluent regulations. [4] The strategy employed to tackle the challenge involved the development of an AI-based neural soft sensor, a neural identification model, and an all-encompassing control strategy for WWTPs. [4] The input variables (i.e., secondary variables) included wastewater concentrations measured through a physical online sensor system. The most suitable input variables for the AI algorithm were determined using the Principal Component Analysis (PCA) technique. [4] The computational detector featured a neural network structure consisting of two hidden layers with neuron counts of 100, 55, 25, and 3. The dataset consisted of 4,200 data points (sampling time $T_s=15$ minutes). The authors employed the neural identification model developed to assess the dynamic behaviour of WWTPs with respect to COD, TN, and TSS. [4] A DNN (Deep Neural Network) model was used to map the temporal evolution and management parameters (influent flow rate and $\text{NH}_4\text{-N}$). One more, Principal component analysis (PCA) for variable selection. The developed framework consisted of a standalone neural network employing a feedforward configuration of 50 – 35 – 15 – 1 neuron was trained using 1,500 data entries recorded at 15-minute intervals ($T_s=15\text{min}$). The outcomes demonstrated that the neurogenetic controller achieved an R^2 value in the range of 0.88 to 0.91, alongside a mean absolute percentage deviation between 2.99% and 4.52%. Compared to the conventional decoupled PID control, the neurogenetic method reduced average costs by 12.89%. [4]

The effectiveness of the treated effluent is influenced by both the characteristics of the incoming wastewater and the operational conditions within the treatment plant. [4] Across various studies, input variables used in the models differed significantly, as multiple factors influence WWTP performance. [6]



A significant portion of research in this domain (around 52%) has drawn upon indicators of influent volume and composition as input variables. This trend highlights a common preference among researchers for applying artificial neural networks (ANNs) to capture the intricate and non-linear interactions between influent characteristics and effluent quality. [4] For instance, in a study by Bekkari & Zeddouri, [10] influent parameters such as pH, temperature, TSS, TKN (Total Kjeldahl Nitrogen, a measure of the amount of nitrogen in an environmental sample), BOD, and COD were used to predict the COD level in treated water from a WWTP in Algeria. [10]

Wastewater treatment monitoring data often involves complex and non-linear chemical relationships. Due to this complexity, ANNs, which are inherently suited for non-linear modelling, can accurately predict pollutant removal in WWTPs. [7]

ANNs are widely utilised in water-related research due to their capacity to learn complex non-linear, multi - input/output relationships from historical data during the training process. They are particularly valuable when there is typically based on physical and chemical principles. [7]

Mechanistic models often depend on material and energy balances and empirical formulae, however, these can be inaccurate and require numerous assumptions to remain manageable. As a result, such models carry significant limitations. [7]

4. ADVANCEMENTS AND THE INCORPORATION OF AI INTO WASTEWATER TREATMENT PRACTICES

Biological WWTPs are highly complex and adaptive systems, shaped by the interplay of microbial activity in the incoming wastewater and the multitude of processes involved. As a result, accurately modelling these systems and evaluating their efficiency presents significant difficulties, which has led to a surge in research interest. Recently, AI methodologies, along with associated tools, have become increasingly valuable for forecasting effluent quality or assessing the efficiency of contaminant removal in both *large-scale* and experimental wastewater treatment systems. [4]

Regulatory requirements and discharge limits for effluent from WWTPs are becoming increasingly stringent. The challenging physical environment and demanding conditions within WWTPs make the transparent model-based approach (a method relying on detailed process understanding) impractical for sensor development. [4]

Artificial intelligence (AI) is increasingly being employed to support the creation of virtual measurement tools, enhance automated control mechanisms, and improve optimisation processes in wastewater processing systems. For example, BOD₅ (Biological Oxygen Demand over five days) represents a key indicator in evaluating the condition of water for specific uses, yet the standard test is both time consuming and costly. AI-based soft sensors are being developed to address these challenges. [4] For instance, Osman and Li (2020) [11] developed a BOD₅ sensor using six input variables, achieving a high R^2 value of 0.99. Furthermore, soft sensors have been created to assist in the control of WWTPs determined in relation to environmental weather patterns, given that such factors can alter both the rate of incoming flow and the concentration of specific constituents. [4]

To improve predictive performance, the study employed genetic algorithms to optimise the quantity of nodes within every concealed layer of the neural network framework. [11] This evolutionary approach enabled the identification of the most suitable network architecture for BOD₅ prediction, thereby enhancing the model's accuracy and generalisation capability. [11]

Feed-forward neural networks operate through a three-layer structure: an input layer containing a vector of pre-processed parameters, an intermediary layer incorporating non-linear transformations, followed by the concluding layer that compares predicted values to actual results. These networks are often trained using the Levenberg-Marquardt algorithm, which enhances convergence and accuracy. Moreover, integrating lagged input data through auto-regressive models helps improve predictive performance in dynamic conditions. [12]

Despite their effectiveness, ANN-based models can be computationally intensive, especially when processing large datasets, making them less feasible for smaller treatment facilities with limited hardware resources. Support vector machines, on the other hand, provide a more efficient alternative, delivering accurate predictions while maintaining lower computational demands. [12]

Adaptive systems based on neural networks and fuzzy logic offer a combined approach, merging the learning ability of neural networks with the flexibility of fuzzy logic to manage uncertainty and variability in biological treatment processes. These models are particularly useful when dealing with sensitive biomass systems, such as aerobic granular sludge reactors, where influent conditions may fluctuate frequently. [12]



Recent studies have also explored hybrid frameworks that combine ANFIS with support vector regression (SVR) in a two-stage modelling process. This approach allows individual prediction of key output parameters and provides greater flexibility in adjusting for errors, leading to improved overall model robustness and accuracy. [12]

5. CONCLUSIONS AND PERSPECTIVES

Based on the current application of AI in wastewater treatment plants, several key aspects can be highlighted in context prediction in AI applications, AI algorithms, soft and online sensors, algorithm comparison, and AI for image analysis. [10]

Prediction in AI applications: Almost all AI applications in WWTPs focus on prediction. These predictions include outputs such as effluent quality parameters (e.g., COD, BOD₅, NH₄-N), energy consumption, influent data, aeration time, sludge bulking, and sludge settleability. [10]

A variety of artificial intelligence approaches, including foundational, combined, and advanced models, along with both supervised and unsupervised techniques, have been evaluated using real-world data. While artificial neural networks (ANN) are the most prevalent algorithm, deep learning (DL) techniques such as DNNs, LSTMs, CNNs, and hybrid approaches are steadily gaining more attention. [4]

Soft sensors are becoming essential in supporting the operation of WWTPs. Their integration into intelligent control systems adds more value than merely using model-based predictive sensors. [10]

Comparisons between algorithms applied to different datasets or studies remain relative. Although R² values and error metrics offer preliminary insight, synchronised data with consistent output units is necessary for producing increased trustworthiness of the findings. [10]

Real-time physical monitoring devices, which produce substantial amounts of information, are likewise utilised in developing AI frameworks based on data patterns. [4] These sensors are particularly useful for supporting WWTP operations, including the detection of faults or anomalies. [4]

Leveraging artificial intelligence for interpreting visual inputs, particularly in scenarios involving identifying aggregated microbial clusters and surface-adhered biological layers or predicting effluent quality, still requires further investigation. [4]

Recommendations for future development of AI in water waste treatment include: data solutions should be developed with a thorough understanding of the processes; *large-scale* data sharing and collaborative research efforts should be encouraged; feature selection techniques it is advisable to integrate dimensionality reduction and feature selection methods-such as principal component analysis, multiple regression techniques, and sequential elimination strategies should be incorporated into AI deployment approaches for WWTPs.; [10] and the integration of online sensors and the data-driven models capable of real-time learning and adaptation represents a promising future direction. [10]

Further exploration into AI-based image analysis for water quality assessment, biofilm detection, and sludge characterisation could significantly enhance the operational efficiency and maintenance of wastewater treatment systems. [3]

The rapid progress in machine learning methodologies opens up diverse possibilities for their use in the wastewater treatment industry. This review specifically explored the role of artificial neural networks in forecasting the effectiveness of WWTPs, focusing on parameters like effluent quality and pollutant removal rates. [9]

A systematic review methodology enabled a focused selection and analysis of studies on artificial neural networks in wastewater treatment, leading to a deeper understanding of model structures and parameter tuning. However, the review did not assess real-world applications of these models. One of the main challenges in applying AI to WWTPs remains the lack of reliable and high-quality data, which is crucial for model accuracy and effectiveness. However, it is crucial to underline that one of the key obstacles to applying these models is the challenge of obtaining reliable and accurate data. [9]

REFERENCES

- [1] A. Capodaglio, “The role of wastewater treatment in achieving sustainable development goals,” *Environmental Science & Ecotechnology*, vol. 6, pp. 100-107, 2021.
- [2] A. Burkov, *The Hundred-Page Machine Learning Book*, 2019.
- [3] D. Z., *Wastewater treatment plant monitoring via a deep learning approach*, 2018.
- [4] T. T. H. N. Q. B. T. X.-T. B. H. H. N. i. D. D. N. X. C. Nguyen, “Artificial intelligence for wastewater treatment,” in *Current Developments in Biotechnology and Bioengineering: Advances in Biological Wastewater Treatment Systems*, 2022, 2022, p. 587-608.



- [5] P. C. S. G. S. K. V. R. Tomar, AI Enabling Tehnologies: A Survey, 2019.
- [6] S.N.P.Rusell, Artificial Intelligence: A modern approach, 2016.
- [7] S. Z. Z, Modeling and Optimalisation of Paper-making Wastewater Treatment Based on Reinforcemen Learning, 2017.
- [8] R. S. Bukhari, "Machine learning: Real-world uses and future challenges," *Tech Horizons Journal*, vol. 2, pp. 1-12, 2021.
- [9] C. C. S. C. O. Marina Salim Dantas, "Artificial neural networks for performance prediction of full-scale wastewater treatment plants: a systematic review," *Water Science & Technology*, vol. 88, no. 6, p. 1447–1470, 2023.
- [10] G. W. K., Deep Learning-based Management For Wastewater Treatment Plants Under Blockmain Environmen, 2020.
- [11] M. S. Zaghloul, "Application of machine learning techniques to model a full-scale wastewater treatment plant with biological nutrient removal," *Journal of Environmental Chemical Engineering*, vol. 10, no. 3, 2022.
- [12] A. Voravich.Ganthavee, "Artificial intelligence and machine learning for the optimization of pharmaceutical wastewater treatment systems," *Environmental Chemistry Letters*, vol. 22, pp. 2293-2318, 2024.

CIP- Каталогизација у публикацији
Народна библиотека Србије

007:004004.8(082)

007:004(082)

004.7(082)

62(082)

INTERNATIONAL Scientific Conference on Information Technology, Computer
Science and Data Science Sinteza (12 ; 2025 ; Novi Sad)

Book of proceedings / International Scientific Conference on Information
Technology, Computer Science and Data Science Sinteza, Novi Sad, May 9 , 2025. ;
[editor-in-chief Milovan Stanišić]. - Belgrade : Singidunum University, 2025 (Belgrade
: Caligraph). - XVI, 566 str. : ilustr. ; 24 cm

"The 12th International Scientific Conference Sinteza was held on May 9, 2025, in
person at the premises of Singidunum University." --> About Sinteza 2025. - Tiraž 200.
- Bibliografija uz svaki rad.

ISBN 978-86-7912-841-6

a) Вештачка интелигенција -- Зборници b) Информационо-комуникационе
технологије -- Зборници v) Рачунарство -- Зборници g) Инжењерство --
Зборници

COBISS.SR-ID 170299913



The 12th International Scientific Conference Sinteza was held on May 9, 2025, in person at the premises of Singidunum University. The conference focused on information technology, computer science, data science, and their applications in engineering systems, education, foreign language teaching, and sports. This year's program placed particular emphasis on topics related to artificial intelligence, machine learning, and data-driven research, especially in the context of solving real-world challenges.

As in previous years, the conference was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia.

Sinteza 2025 once again brought together researchers from Serbia and abroad. A total of 103 papers were submitted, out of which 78 were accepted and 25 were rejected. The conference featured 200 unique authors from 14 different countries, including Serbia. All accepted papers for the Sinteza 2025 conference are scientific papers. Additionally, all accepted papers underwent a thorough review process, including technical, linguistic, and expert evaluation, as well as iThenticate plagiarism checks and verification to assess whether, and to what extent, the text was generated by artificial intelligence tools.

At Sinteza 2025, we welcome academics, researchers, and industry professionals to share their knowledge and insights on the latest developments in Information Technology, Computer Science, and Data Science. We also recognise the importance of supporting the next generation of professionals in these fields, and we are pleased to offer a specially reserved student session at the conference. Each parallel session was interactive and engaging, providing presenters the opportunity to share their research papers, case studies, and innovative projects. Conference participants actively engaged in discussions, offering feedback and exchanging insights with experts in the field.

