

University of Novi Sad
Technical faculty "Mihajlo Pupin" Zrenjanin

**Proceedings of the 14th International Conference
on Applied Internet and Information Technologies
AIIT 2024**

8 November, 2024, Zrenjanin, Serbia





University of Novi Sad
Technical faculty
"Mihajlo Pupin"
Zrenjanin
Republic of Serbia



XIV INTERNATIONAL CONFERENCE ON APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2024
PROCEEDINGS



November 8, 2024
Zrenjanin
Serbia

Proceedings publisher and organizer of the conference:

University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Republic of Serbia

For publisher:

Milan Nikolic, PhD

Dean of Technical faculty "Mihajlo Pupin",
Zrenjanin, Republic of Serbia

Proceedings editors:

Ivana Berkovic, PhD

Kostandina Veljanovska, PhD

Conference Chairmans:

Ivana Berkovic, Technical faculty "Mihajlo Pupin", University of Novi Sad, Zrenjanin, Serbia

Kostandina Veljanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Technical preparation of the proceedings:

Zoltan Kazi, PhD

Sinisa Mihajlovic, MSc

Aleksandra Stojkov Loncarski, MSc

Milica Mazalica, MSc

Cover design:

Visnja Ognjenovic, PhD

e-Proceedings

ISBN 978-86-7672-379-9

Disclaimer:

All rights reserved. No part of this proceeding may be reproduced in any form without written permission from the publisher. The publisher and editors are not responsible either for the statements made or for the opinion expressed in this publication. The authors solely are responsible for the content of the papers and any copyrights, which are related to the content of the papers.

CIP - Каталогизacija y publikaciji
Библиотеке Матице српске, Нови Сад

004(082)(0.034.4)

INTERNATIONAL Conference on Applied Internet and Information Technologies (14 ; 2024 ; Zrenjanin)

Proceedings [Elektronski izvor] / XIV International Conference on Applied Internet and Information Technologies AIIT 2024, Zrenjanin, November 8th, 2024 ; [organizer] Technical Faculty "Mihajlo Pupin", Zrenjanin ; [editors Ivana Berkovic, Kostandina Veljanovska]. - Zrenjanin : Tehnički fakultet "Mihajlo Pupin", 2024. - 1 elektronski optički disk (CD-ROM) : tekst, ilustr. ; 12 cm

Sistemska zahtevi: Nisu navedeni. - Naslov sa nasl. ekrana. - Elektronska publikacija u formatu pdf opsega 504 str. - Bibliografija uz svaki rad. - Registar.

ISBN 978-86-7672-379-9

a) Информационе технологије -- Зборници

COBISS.SR-ID 158823945

Introduction

International Conference on Applied Internet and Information Technologies is an annual Conference that we started in 2012 after successful results of the International Conference on Information and Communication Technologies for Small and Medium Enterprises in 2011. This year, the fourteenth Conference is hosted in Zrenjanin and presents collaboration between the University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia and the University "St. Kliment Ohridski", Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia.

The Conference was financially supported by the Provincial Secretariat for Higher Education and Scientific Research, Novi Sad. The Technical Faculty "Mihajlo Pupin" has provided the necessary technical support.

This year we had gathered our colleagues, scientists, researchers and students from 13 countries: Canada, USA, India, Russia, Germany, Denmark, Georgia, Slovakia, Hungary, Croatia, Bosnia and Herzegovina, North Macedonia and Serbia. They presented papers and promote the results of research and scientific work in the field of information technology. Out of more submitted papers, 66 were selected for presentation on Conference and publishing in Proceedings.

Some of Conference topics are: Artificial Intelligence, Intelligent Systems, Data Science, Big Data Technologies, Business Intelligence, IT Support to Decision-making, Information Systems, Software Engineering and Applications, Communications and Computer Networks, Data and System Security, Computer Graphics and Visualization, IT Management, etc.

During the conference were organized in six sessions: two plenary sessions, two Oral Sessions and two Poster sessions. The introductory lectures were:

- *Text Mining*, held by Slavica O'Connor, Canada,
- *IT equipment and software for training, modeling and data analysis for flood and forest fire prevention, protection and management in project SOLVE*, held by Andrijana Bocevska, North Macedonia,
- *Reducing Manual Labeling Effort by Identifying the Most Informative Unlabeled Data via Active Learning*, held by Velibor Ilić, Serbia,
- *Digital platform for monitoring and forecasting the environmental situation of the Baikal natural territory*, held by Igor V. Bychkov, Russia.

The AIIT program committee would like to thank the authors of the papers for their contribution. All submitted papers were peer-reviewed by members of the committee and the other eminent reviewers. All submitted papers were peer-reviewed through the double-blind review process. Also, the AIIT program committee would like to express special gratitude to the reviewers for their tremendous work done for selecting the papers with their valuable comments and suggestions that contributed to improve the quality of the papers.

AIIT 2024 was very successful conference with fruitful exchange of experiences among the participants and contribution to the further development of Internet and Information technologies research.

Next year the conference will be held in Bitola.

Conference chairs:

Ivana Berković, University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia (chair)

Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia (co-chair)



Technical Faculty "Mihajlo Pupin" Zrenjanin
University of Novi Sad
SERBIA
<http://www.tfzr.uns.ac.rs/>

ORGANIZATION PARTNERS:



Faculty of Information and Communication Technologies -
Bitola
"St. Kliment Ohridski" University - Bitola
NORTH MACEDONIA
<http://fikt.uklo.edu.mk/>



Matrosov Institute for System Dynamics and Control Theory of
Siberian Branch of Russian Academy of Sciences, Irkutsk,
RUSSIA
<http://idstu.irk.ru/>



Irkutsk State Transport University (IrGUPS)
Irkutsk, RUSSIA
<https://www.irgups.ru/>



Faculty of Engineering
South-west university "Neophyte Rilsky"-Blagoevgrad
BULGARIA
<http://www.swu.bg/>

Conference Chairs Biographies

Ivana Berković, University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia (chair)

Ivana Berković, Ph.D. is a Full Professor at the Technical Faculty "Mihajlo Pupin" in Zrenjanin, Serbia. She obtained her Bachelor's degree from the Faculty of Sciences in Novi Sad and completed her Master's and Ph.D. degrees at the Technical Faculty "Mihajlo Pupin" in Zrenjanin, specializing in Logic Programming and Automated Reasoning. Since 1987, Prof. Berković has been teaching at the Technical Faculty "Mihajlo Pupin." In 2008, she was appointed as a Full Professor at the University of Novi Sad. She was associated dean from 2002 to 2010. She is a member of Professional Councils of the University of Novi Sad. Her research interests include Artificial Intelligence, Automated Reasoning, Logic Programming Languages, and Computer Graphics. She has authored numerous scientific papers, textbooks, and software products. She participated as researcher or leader of 12 national funded projects. She is a member of editorial board of Journal ComSIS and member of program and organization committee of several international conferences.

Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (co-chair)

Kostandina Veljanovska, Ph.D. completed her education at the University "Sts. Kiril i Metodi", Skopje (BSc in Computer Science), at the University of Toronto, Toronto (MSc in Applied Engineering) and got her MSc and also her PhD in Technical Sciences at the University "St. Kliment Ohridski", Bitola, R. N. Macedonia. She has completed postdoc in Artificial Intelligence at the Laboratory of Informatics, Robotics and Microelectronics at the University of Montpellier, Montpellier, France. She worked as a Research Assistant at the Faculty of Applied Science, University of Toronto, Canada. She also, worked as a researcher at research team for Constraints, Learning and Agents at LIRMM, University of Montpellier. Currently, she works as a Full Professor in Artificial Intelligence and Systems, Computer Science and Computer Engineering at the Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" -Bitola, Republic of N. Macedonia. Since 2022 she is vice-dean for Science and Collaboration. Her research work is focused on artificial intelligence, machine learning techniques, intelligent systems and human – computer interaction. She has published numerous scientific papers in the area of interest. She is a reviewing referee for several publishing houses, journals with significant impact factor in science and also, member of editorial board of several international conferences.

Organizing Committee

Ivana Berković, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia, president

Višnja Ognjenović, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Dalibor Dobrilović, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Biljana Radulović, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Željko Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Ljubica Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Eleonora Brtko, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Vladimir Brtko, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Zoltan Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Siniša Mihajlović, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Velibor Premčevski, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Aleksandra Stojkov Lončarski, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Maja Gaborov, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Milica Mazalica, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Igor Vecštejn, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Marko Blažić, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Dalibor Šeljmeši, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Jovana Borovina, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Vuk Amižić, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Dilan Dobardžić, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Bojan Vujanov, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Vladimir Karuović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Kostandina Veljanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Blagoj Ristevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Snežana Savoska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Monika Markovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Zoran Kotevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Božidar Milenkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Nikola Rendeovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Andriana Bocevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Tome Dimovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Natasa Blažeska-Tabakovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Mimoza Bogdanoska-Jovanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Marina Blažeković Toshevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Program Committee

Ivana Berković, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia, president

Željko Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Ljubica Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Višnja Ognjenović, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Dalibor Dobrilović, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Dragica Radosav, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Dragana Glušac, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Biljana Radulović, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Vladimir Brtko, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Zoltan Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Eleonora Brtko, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Jelena Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Vesna Makitan, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Nadežda Ljubojev, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Ćočkalo Dragan, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Kavalić Mila, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Terek Stojanović Edit, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Prvulović Slavica, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Desnica Eleonora, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Palinkaš Ivan, Technical faculty "Mihajlo Pupin", Zrenjanin, University of Novi Sad, Serbia

Igor Nedelkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Makedonia

Blagoj Ristevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Makedonia

Aleksandar Markovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Makedonia

Violeta Manevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Pece Mitrevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Ilija Jolevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Dragan Gruevski, Faculty of Information and Communication Technologies, University "St. Kliment

Ohridski", Bitola, North Macedonia

Kostandina Veljanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Monika Markovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Snežana Savoska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Ramona Markoska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Mimoza Bogdanoska-Jovanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Nataša Blažeska Tabakovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Božidar Milenkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Zoran Kotevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Andrijana Bocevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Sonja Mančevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Marina Blažekovik Toševski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Nikola Rendeovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Lela Ivanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Ilija Hristoski, Faculty of Economics - Prilep, Macedonia, University "St. Kliment Ohridski", Bitola, North Macedonia

Elena Vlahu Gjorgievska, University of Wollongong, Australia

Igor Bychkov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

Evgeniy Cherkashin, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

Alexander Feoktistov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

Roman Kostromin, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

Andrey Gachenko, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia

Andrey Mikhailov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia

Anastasia Popova, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia

Alexey Daneev, Irkutsk State Transport University, Irkutsk, Russia

Denis Sidorov, Melentiev Energy Systems Institute of Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

Viacheslav Paramonov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

Andrey Dorofeev, Institute of High Technologies, Irkutsk National Research Technical University, Irkutsk, Russia

Branko Markoski, Faculty of Technical Sciences, University of Novi Sad, Serbia

Dinu Dragan, Faculty of Technical Sciences, University of Novi Sad, Serbia

Srdjan Popov, Faculty of Technical Sciences, University of Novi Sad, Serbia

Gogolák László, Subotica Tech - College of Applied Sciences, Subotica, Serbia

Zlatko Čović, Subotica Tech - College of Applied Sciences, Department of Informatics, Subotica, Serbia

Nataša Gospić, Faculty of transport and traffic engineering, Belgrade, Serbia

Željko Trpovski, Faculty of Technical Sciences, University of Novi Sad, Serbia

Željko Eremić, College of Technical Sciences - Zrenjanin, Serbia

Rajnai Zoltán, Obuda University, Budapest, Hungary

Tünde Anna Kovács, PhD, Óbuda University, Hungary

Zoltán Nyikes, PhD, Milton Friedman University, Hungary

Mirjana Pejic Bach, Faculty of Economics, University of Zagreb, Croatia

Androklis Mavridis, Aristotel University of Thessaloniki, Greece

Madhusudan Bhatt, R.D. National College, University of Mumbai, India

Amar Kansara, Parth Systems LTD, Navsari, Gujarat, India

Narendra Chotaliya, H. & H.B. Kotak Institute of Science, Rajkot, Gujarat, India

Marijana Brtko, Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, São Paulo, Brazil

Slobodan Lubura, Faculty of electrical engineering, University of East Sarajevo, Bosnia and Hercegovina

Edit Boral, Johns Hopkins University Center for Talented Youth, New York, USA

Dana Petcu, West University of Timisoara, Romania

Marius Marcu, "Politehnica" University of Timisoara, Romania

Filip Tsvetanov, South-west University "Neophyte Rilsky", Faculty of Engineering, Blagoevgrad, Bulgaria

Francesco Flammini, School of Innovation, Design and Engineering, Division of Product Realisation, Mälardalen University, Eskilstuna, Sweden

Deepak Chahal, Jagan Institute of Management Studies (JIMS, Rohini Sector-5), New Delhi, India

Abdel-Badeeh M. Salem, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

Dragan Peraković, University of Zagreb, Faculty of Transport and Traffic Sciences, Croatia

Gordana Jotanović, University of East Sarajevo, Faculty of Transport and Traffic Engineering, Doboj, Bosnia and Herzegovina

Goran Jauševac, University of East Sarajevo, Faculty of Transport and Traffic Engineering, Doboj, Bosnia and Herzegovina

Ming Chen, Zhejiang University, China

Ertuğrul AKBAŞ, Esenyurt University, Istanbul, Turkey

Gururaj Harinahalli Lokesh, Department of IT, Manipal Institute of Technology, Bengaluru, India

TABLE OF CONTENTS

PLENARY PAPERS	1
Text Mining <i>Slavica O'Connor</i>	2
IT equipment and software for training, modeling and data analysis for flood and forest fire prevention, protection and management in project SOLVE <i>Andrijana Bocevska, Igor Nedelkovski, Aleksandar Markoski, Zoran Kotevski, Kostandina Veljanovska, Blagoj Risteovski, Snezana Savoska</i>	17
Digital platform for monitoring and forecasting the environmental situation of the Baikal natural territory <i>Igor V. Bychkov, Evgeniy S. Fereferov, Roman K. Fedorov</i>	28
Reducing Manual Labeling Effort by Identifying the Most Informative Unlabeled Data via Active Learning <i>Velibor Ilic</i>	35
REGULAR PAPERS	44
Malware and Social Engineering Treats <i>Aleksandar Bulajic</i>	45
Investigating the impact of social networks on viral marketing <i>Magdalena Ninkov, Snežana Jokić, Đorđe Alavuk, Nikola Jančev, Marjana Pardanjac</i>	54
Artificial Intelligence, Privacy, and Information Security in Critical Infrastructure Systems <i>Rade Dragović, Nadežda Ljubojev, Dalibor Dobrilović, Bojan Perović, Vladimir Milošev, Milada Novaković</i>	64
Programming model for GenICam-based cameras <i>Bojan Torbica, Dragan Ivetić, Aleksandar Kupusinac</i>	72
Security of Information Systems in Solar Energy Overview of Approaches and Practices <i>Luka Djordjević, Borivoj Novaković, Mihalj Bakator, Velibor Premčevski, Stefan Ugrinov</i>	79
Could Facebook's user interface be more usable for seniors? <i>Dragana Bodiroga, Dragan Ivetić</i>	85
Design and Development of Laser Engraving Machine <i>Evgenija Divitarova, Andrijana Bocevska</i>	93
Quality of UML Class Diagram: An Overview <i>Amižić Vuk, Kazi Zoltan</i>	100
On the Energy Consumption of Global Video Streaming <i>Zoran Kotevski, Igor Nedelkovski, Aleksandar Markoski, Andrijana Bocevska</i>	106

The Social Network Facebook as a Marketing Tool in the IT Retail Sector <i>Mitar Jovanović, Olivera Stupar, Mila Kavalić, Sanja Stanisavljev, Verica Gluvakov, Snežana Mirković</i>	114
The impact of Artificial Intelligence on personalization strategies in digital marketing <i>Verica Gluvakov, Mila Kavalić, Mihalj Bakator, Stefan Ugrinov, Sanja Stanisavljev, Snežana Mirković</i>	121
Algorithmic Management in the Platform Economy <i>Ivana Simić, Lucija Stojković</i>	127
Quarterly Trends in Soil Moisture And Air Temperature: A Python Based Analysis <i>Milena Zeljković, Dinu Dragan, Srđan Popov, Tanja Vranić</i>	132
A Petri Net-Based Approach to Modeling Concurrency and Parallelism in Computer Architectures <i>Ilija Hristoski, Jelena Stojanov</i>	140
Comparative Analysis: Web, Native, and Progressive Web Applications <i>Sara Poparić, Dinu Dragan, Dušan B. Gajić, Veljko B. Petrović</i>	148
Testing the capabilities of artificial intelligence and prompt engineering in the field of agrochemical soil analysis with fertilization recommendations <i>Željko Eremić, Milana Drašković</i>	156
A comparative study of Software Development Life Cycle (SDLC) models <i>Buen Bajrami, Ilija Jolevski, Kostandina Veljanovska</i>	162
Program Model for a Visual Editor of Gamepad Haptic Effects <i>Vasilije Bursać, Dragan Ivetić, Aleksandar Kupusinac</i>	173
Upgrading Traditional E-Commerce Systems With A Knowledge-Based Recommendation System <i>Teodora Siljanoska, Natasa Blazeska Tabakovska</i>	181
Internet of Things in the Automotive Industry: A Review <i>Maša Knežević, Goran Sladić, Branko Markoski, Aleksandra Stojkov Lončarski</i>	187
The role of augmented and virtual reality in improving consumermarketing experiences <i>Mihalj Bakator, Luka Đorđević, Borivoj Novaković, Mića Đurđev, Stefan Ugrinov</i>	193
Investigating IT Experts' Attitudes Toward the Frequency of Agile Meetings: A Thematic Analysis <i>Maja Gaborov, Zeljko Stojanov, Srdjan Popov, Dragana Kovac, Igor Vecstejn, Gordana Jotanovic, Goran Jausevac</i>	200
Pricing Dynamics in the Digital Housing Market: An Analytical Comparison of RentLingo.com and RentDigs.com <i>Gisele Habibulla</i>	208

From Digital Substations to Cybersecurity: Protecting the Evolving Power Transmission Network <i>Mugdin Agić, Mia Lešić Aganović, Emir Skejić</i>	215
Machine learning and theory of information in natural language processing <i>Zlatko Radovanovic, Dejan Djukic</i>	223
Artificial intelligence in electrical engineering <i>Dejan Djukic, Stefan Popovic</i>	231
A multi-task management system based on PHP and MySQL <i>Buen Bajrami, Violeta Manevska, Kostandina Veljanovska</i>	238
Proactive E-Services in Preventive Health Care Utilizing Knowledge Management <i>Mimoza Bogdanoska Jovanovska, Marija Petrevska, Natasa Blazeska-Tabakovska, Snezana Savoska</i>	247
Cyber attacks and defense in the context of the application of regulations in the Republic of Serbia – challenges and proposals for improving the security of ICT systems <i>Tamara Milic, Jelena Bjelica</i>	255
Challenges in Evaluating Large Language Models: Insights from Five Critical Benchmarks <i>Dilan Dobardžić, Višnja Ognjenović, Vladimir Brtko, Jelena Stojanov, Katarina Vignjević</i>	262
Retrieval-Augmented Generation (RAG) in Local Large Language Models (LLMs) <i>Dilan Dobardžić</i>	267
Optimizing Renal Scintigraphy Interpretation with an Interactive Deep Learning Tool <i>Olivera Lazić, Marijana Matkovski, Dunja Vrbaški, Ana Jakovljević, Dragan Burić</i>	272
Techniques of Software Integration Testing <i>Zoltan Kazi, Ljubica Kazi, Madhusudan Bhatt, Narendra Chotaliya</i>	280
AI at the Edge: Trends and Innovations in Tiny Machine Learning Models for IoT and Embedded Systems <i>Aneta Trajkovska, Aleksandar Markoski</i>	287
Using predictive analytics and machine learning in Python to analyze and predict student success <i>Katarina Vignjević, Marko Blažić, Biljana Radulović, Milan Marković, Uroš Šarenac, Dilan Dobardžić</i>	295
An Overview of Multimedia - audio data visualization processing using Business intelligence tools <i>Marko Blažić, Srđan Popov, Višnja Ognjenović, Katarina Vignjević, Milan Marković, Slobodan Nadrljanski</i>	302
The Influence of the Trademark on the Visual Identity of Social Networks Through User Experience <i>Nikola Jerković, Ivana Berković, Višnja Ognjenović</i>	310

The Inheritance Relationship in the Teaching of Object-Oriented Programming <i>Aleksandar Kupusinac</i>	316
Testing of the e-security systems in the automotive industry <i>Predrag Novokmet, Vesna Makitan, Dragana Glušac, Eleonora Brtko, Nikola Jovanov</i>	320
Review of Different Types of Computer-Aided Engineering <i>Raul Turmanidze, Predrag Dašić, Vazha Shilakadze, Giorgi Popkhadze</i>	326
Communication Challenges in Agile Meetings <i>Maja Gaborov, Miodrag Kovačević, Nemanja Tasić, Dragan Kreculj, Nada Ratković Kovačević, Dragana Kovač</i>	334
Analysis of GPT-4o Application in Solving Simple Static Beams Based on Uploaded Images: Potentials and Limitations <i>Ljubica Lazić Vulićević, Željko Eremić, Valentina Mladenović</i>	341
Six-Layered Application of Artificial Intelligence in Business Intelligence <i>Veselina Naneva, Kremena Stefanova</i>	347
Personal Data Protection in eGovernment <i>Rade Dragović, Nadežda Ljubojev, Dalibor Dobrilović, Bojan Perović, Milenko Đuričić</i>	355
Encryption as an Information Security Function in Critical Infrastructure Systems <i>Vladimir Kačanovski, Rade Dragović, Dragan Dragović, Bojan Perović</i>	363
Interactive faculty map and timetable web application <i>Ana Marija Ognjenović, Bosko Nikolic, Visnja Ognjenovic</i>	372
Development of a Responsive Web Application to Support the Creation of Class Schedules <i>Ivana Samardzic, Dragoljub Krneta</i>	380
Experience with Proof-of-Concept Applications with a Micro-frontend Architecture: The Vet Clinic Box Monitoring App Case <i>Katarína Semanová, Csaba Szabó</i>	388
The Perception of Risks in Digital Payment Services <i>Nikola Jerković, Ana Jovanović, Ivana Berković, Edit Boral</i>	396
Neural networks in pellet combustion control - research report for OZONE 55 <i>Stefan Popovic, Sonja Djukic Popovic, Dejan Djukic, Nebojsa Denic</i>	402
Revolutionizing education through adaptive technologies <i>Nemanja Tasić, Dragana Glušac, Vesna Makitan, Miodrag Kovačević, Tamara Milić</i>	408
The Role of Emerging Technologies in Transforming Supply Chains <i>Stefan Ugrinov, Edit Terek Stojanović, Dragana Kovač</i>	413
Regression Gradient Initialization: A New Approach to Weight and Bias Initialization <i>Dalibor Šeljmeši, Jelena Stojanov, Vladimir Brtko, Dragan Peraković</i>	420

KNN Algorithm Implementation in Real-World Problem of Water Quality Classification <i>Kostandina Veljanovska, Aneta Trajkovska, Nikolce Veljanovski</i>	430
Solar-Powered Sensor Station's Energy Consumption Simulator <i>Dalibor Dobrilović, Jasmina Pekez, Eleonora Desnica, Ivan Palinkaš</i>	436
Implementing Master-Detail Interface Design in an MVC .NET Web Application for Gym Membership Management <i>Nikola Jovanov, Eleonora Brtka, Predrag Novokmet, Milica Mazalica, Aleksandra Stojkov Loncarski</i>	443
Adaptive Biometric Systems: Challenges and Opportunities <i>Anita Petreska, Blagoj Ristevski, Ivica Dimitriovski, Saso Nikolovski, Daniela Slavkovska</i>	451
Using Machine Learning Algorithms of Stroke Prediction <i>Daniela Slavkovska, Anita Petreska, Blagoj Ristevski, Saso Nikolovski, Nikola Rendevski</i>	461
Evaluation of the Efficiency of User Interfaces for AI Tools <i>Gordana Jotanovic, Goran Jausevac, Vladimir Brtka, Armin Podanovic, Jelena Simic, Ajla Nurkanovic</i>	472
Integration of Image Recognition Systems in Web Applications <i>Nikola Jovanov, Eleonora Brtka, Vesna Makitan, Predrag Novokmet, Albina Kószó</i>	480
Cloud services modeling for long-term intellectual capitalprotection <i>Saso Nikolovski, Bozidar Milenkovski, Anita Petreska, Daniela Slavkovska</i>	488
Object Tracking in Video Stream <i>Marko Beljin, Ivan Gasic, Zeljko Stojanov, Peter Odry, Zoltan Vizvari, Vladimir Tadic</i>	496
INDEX OF AUTHORS	502
COMPANY PRESENTATION	504

AIIT 2024

PLENARY PAPERS



Text Mining

Slavica O'Connor¹

¹ Seneca College of Applied Arts and Technology, 1750 Finch Avenue East, Toronto, Ontario, M2J 2X5, Canada

slavica.oconnor@senecapolytechnic.ca, oconnorsla@gmail.com

Abstract:

Organizations gather huge volumes of data, from a variety of sources, on a daily basis, and deal with the problem of how to make valuable use of it. There has been a need to expand the scope of business intelligence to include text information from diverse documents. Text mining is an exciting research area that tries to make beneficial use of such data and solve the information overload problem. This paper provided an overview of text mining, and its main techniques, information retrieval, information extraction, text classification, text clustering, and core mining operations. Major limitations of text mining result from conceptual issues in computational linguistics. Nonetheless, the brisk development of text mining tools is expected to address some of these limitations in the foreseeable future.

Keywords:

Text Mining, Data Mining, Machine Learning, Information Retrieval, Information Extraction, Data Science, Business Intelligence

1. Introduction

The emerging developments in computer and network technologies, as well as the strong presence of the electronic commerce environment, have contributed to the vast amount of online text data that continues to grow rapidly. Text plays a significant role in providing decision makers with an expanded model of business intelligence than in the past. Business intelligence is not limited to numeric reports and graphs. Decision makers extensively benefit from effective and proper use of information available in text format. Text mining techniques are fundamental for dealing with the enormous quantity of text that is available to us today. Text mining is looking for bits of information that can be logically grouped in order to enable users to derive conclusions that had not been previously made.

1.1. Limitations

This paper is not a comprehensive review of the entire field; rather, only fundamental concepts and basic areas have been selected.

1.2. Definition of Terms

Table 1:
Terms and Definitions

Term	Definition
Text Mining	Marty Hearst, a text data mining researcher, defines text mining as “the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources” [1].
Data Mining	A process of selection, exploration, and modeling of very large data sets, in order to discover previously unsuspected, predictive and descriptive patterns and establish relationships.

Information Extraction	A subfield of text mining that extracts information from text documents into a structured data and attempts to move text mining onto the same ground with the structured world of data mining.
Information Retrieval	A computer science discipline which studies the retrieval of information from a collection of written documents.
Natural Language Processing	A subfield of artificial intelligence and computational linguistics dedicated to converting information expressed in spoken and written human languages into computer language via specialized software.
Computational Linguistics	Part of the science of language that uses computation as an investigative tool.
Ontology	Ontology is a formal, explicit, shared conceptualization of a domain, intended to facilitate semantic interoperability among intelligent information systems where diverse software components, computing devices, knowledge, and data, are involved [2].
Taxonomy	Taxonomy is the science of classification according to a pre-determined system, with the resulting catalogue used to provide a conceptual framework for discussion, analysis, or information retrieval. In practice, a good taxonomy should be simple, easy to remember, and easy to use [3].
Data Science	Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
Data Analytics	Data analytics is the collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision-making.
Data Analysis	Data analysis is the practice of working with data to glean useful information, which can then be used to make informed decisions.
Data Analytics vs Data Analysis	Data analytics has a broader scope and encompasses the entire data lifecycle (collection, cleaning, organization). The scope of data analysis is narrower as it focuses on the in-depth examination and interpretation of prepared data. It focuses on deriving insights for decision-making.
Machine Learning	Machine learning is a subfield of artificial intelligence that uses algorithms trained on data sets to create models that enable machines to perform tasks that would otherwise only be possible for humans, such as categorizing images, analyzing data, or predicting price fluctuations.
Deep Learning	Deep learning is a subset of machine learning that uses artificial neural networks to mimic the learning process of the human brain.
Artificial Intelligence	Artificial intelligence (AI) is a set of technologies that enable computers to perform a variety of advanced functions, including the ability to see, understand and translate spoken and written language, analyze data, make recommendations, and more.
Decision Support System	A decision support system (DSS) is an information system that supports business or organizational decision-making activities.
Business Intelligence (BI)	A wide category of methods, technologies and software for collecting, storing, structuring, integrating, analysing, accessing, and presenting the data with the objective to help enterprise users make better business decisions.

Data Warehouse	Bill Inmon, considered as the father of data warehousing, has defined data warehousing as “a database containing subject oriented, integrated, time variant and non-volatile information used to support the decision making process.” Another well-known data warehouse theorist and practitioner, Ralph Kimball, describes a data warehouse as “a copy of transaction data specifically structured for query and analysis” [4].
Document Warehouse	Dan Sullivan states that a document warehouse “is a repository of textual information that is categorized and organized in such a way as to integrate semantically related texts so that they are accessible to end users and provide relatively high levels of information retrieval precision and recall in a decision-support environment” [6].
Data Lake	A data lake is a centralized repository that ingests, stores, and allows for processing of large volumes of data in its original form. It can accommodate all types of data, which is then used to power big data analytics, machine learning, and other forms of intelligent action.

2. Text Mining and Related Techniques

In this section, literature related to this research will be reviewed, including text mining definition, architecture, information retrieval, information extraction, text categorization, text clustering, and data mining. The reader will be presented with an overview of associated disciplines and how text mining relates to, and differs from, those. The aim of this section is to establish a theoretical base for exploring text mining tools and provide understanding of what text mining tools can offer.

2.1. Text Mining Definition and Architecture

From researching textbooks and online resources, several definitions of text mining can be found. The most extensive definition identifies text mining as any operation related to the gathering and analyzing of text, from external sources, for business intelligence purposes [6]. Similarly, Feldman and Sanger [7] broadly define text mining as a knowledge intensive process in which a user interacts with a document collection over time, using a suite of analysis tools. Hearst defines text mining as the discovery of new and previously unknown information, using a computer to automatically extract that information from different written resources [1].

Text mining is a multi-disciplinary field that involves data mining, AI, statistics, machine learning, and computational linguistics to discover patterns, trends, and knowledge from large volumes of unstructured text [13].

System architecture for generic text mining system [7], is shown in Figure 1.

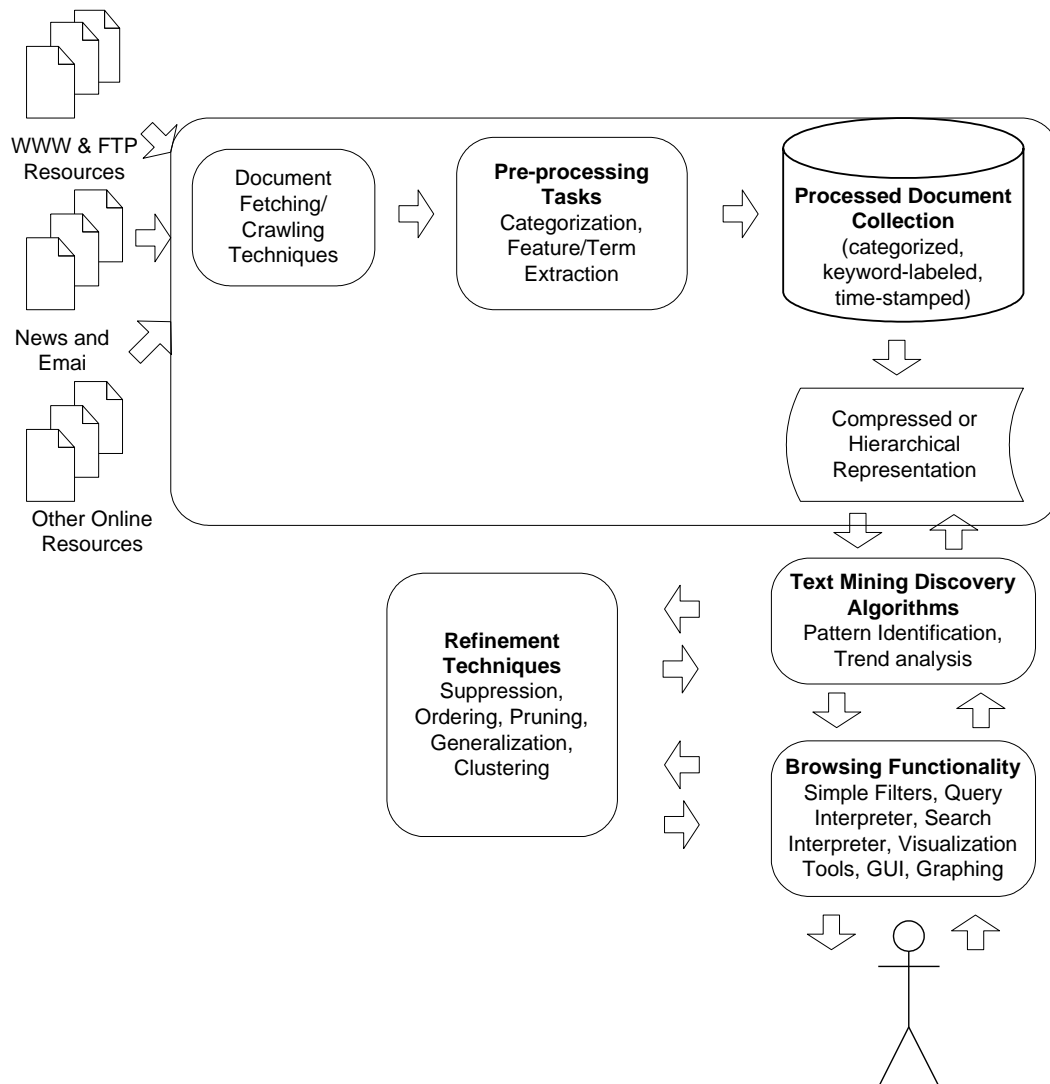


Figure 1: System Architecture for Generic Text Mining System. Adapted from [7].

The generic text mining system architecture includes four main areas [7]:

- *Pre-processing Tasks* – includes necessary processes to prepare data for knowledge discovery operations. Pre-processing tasks create a new collection of documents represented by concepts from the original unstructured data source. Such new document collections are ready for core mining operations. Relevant techniques in this area are information retrieval, text categorization, information extraction, and feature/term extraction.
- *Core Mining Operations* – include pattern discovery, trend analysis, and incremental knowledge discovery. Advanced or specific text mining systems normally enhance core mining operations by leveraging background sources. Such enhanced core mining operations are referred to as *knowledge distillation* processes.
- *Presentation Layer Components and Browsing Functionality* – include GUI, visualization tools, pattern browsing functionality, filters, search interpreters, tools for creating annotated profiles, query language, and query editors.
- *Post-processing or Refinement Techniques* – include methods for suppression, ordering, pruning, generalization, and clustering approaches intended for discovery optimization.

The remaining components in this section explain some of the main techniques, operations and technologies that are part of the text mining architecture.

2.2. Information Retrieval

Information Retrieval (IR) is a broad discipline with many subject areas. Its primary task is to retrieve relevant documents about a particular set of topics in response to a user query. The main steps for information retrieval include query specification, document collection search and return of relevant documents as presented in Figure 2 [8].

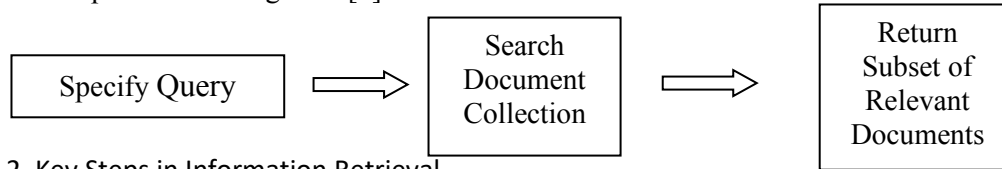


Figure 2. Key Steps in Information Retrieval.

Information Retrieval is dedicated to finding relevant documents, not just simple matches to key words, in a query within an acceptable response time. In reality, the set of retrieved documents differs from the set of relevant documents. There are two ratios for measuring effectiveness: precision and recall [9]. Precision is the ratio representing the number of relevant documents retrieved to the total number retrieved. Recall is the ratio representing the number of relevant documents retrieved to the total number of relevant documents in the collection that are believed to be relevant. In the ideal scenario both ratios have value 1. There are numerous information retrieval models, techniques, utilities, and processing strategies. The mission to discover efficient and effective information retrieval algorithms continues.

2.2.1. Retrieval Strategies

Retrieval strategies address the problem of assigning a degree of similarity between a query and a document. The more terms from the query that a document has, the more it is similar to the query.

“A retrieval strategy is an algorithm that takes a query Q and a set of documents D_1, D_2, \dots, D_n and identifies the Similarity Coefficient $SC(Q, D_i)$ for each of the documents $1 \leq i \leq n$.” [9].

The most common retrieval strategies are [9].

- *Vector Space Model* – This model addresses the following three problems in information retrieval: representing documents, representing queries, and finding documents that meet the criteria of a query. The vector space model is used to solve the first two problems - each document and the query are represented as vectors. The solution to the third problem is to calculate a measure of similarity, the similarity coefficient (SC), between the query and a document vector. SC can be calculated as a dot product of two vectors, but the most commonly used mean is the cosine measure and its variations.
- *Probabilistic Retrieval Model* – For each term in a document collection, a probability is calculated based on the possibility that a term will appear in a relevant document. Then, the similarity measure for terms that match between a document and the query is calculated as the combination of the probabilities for each of the matching terms. Probability theory is used to solve the relevance ranking problem using two major approaches. The first one relies on usage patterns to predict relevance, while the second one uses each term in the query as hints to determine whether a document is relevant or not.
- *Language Model* – This model is based on the idea that documents can be ranked on their possibility to generate the query.
- *Interface Networks* – This model is an enhanced version of probabilistic retrieval strategy that includes additional evidence that a document may be relevant to a query. The strength of this interface is used as the similarity coefficient, and a Bayesian network is used to gather the relevance of a document to a query.
- *Boolean Indexing* – This strategy assigns term weights to each of the terms in the query and to the terms in the document. The weights of the terms are incorporated into a document ranking.

- *Latent Semantic Indexing* – This model uses a matrix to represent the occurrence of terms in documents. The matrix is reduced by using singular value decomposition (SVD) to filter out the noise found. That way, two documents that have the same semantics are located close to each other in a multidimensional space. This model also eliminates major drawbacks associated with the vector space model and is considered one of the most powerful.
- *Neural Networks* – This model consists of nodes and links. Each node has input and output values. Nodes are activated when a query triggers links to documents. The weight of each link is passed on to the document and collected for calculation of a similarity coefficient between the query and the document.
- *Genetic Algorithms* – This model generates an optimal query to find relevant documents by evolution and heredity. An initial query is used with random or estimated term weights. New queries are generated by modifying these weights. New queries that are close to known relevant documents carry on and queries that are far are removed from consequent generations.
- *Fuzzy Set Retrieval* – In this model, each document is mapped to a fuzzy set that contains elements and numbers associated with each element indicating the strength of membership. Boolean queries are mapped into fuzzy set operations (intersection, union, and compliment) that result in strength of membership associated with each document that is relevant to a query. This strength is used as a similarity coefficient.

There is no answer as to which strategy is the best [9]. The evaluation of retrieval strategies remains an area for further research and debate.

2.2.2. Retrieval Utilities

There are many different retrieval utilities that can operate with any retrieval strategy and improve the results of a retrieval strategy. A brief overview of some of the utilities is as follows [9]:

- *Relevance Feedback* – The information retrieval is implemented in multiple passes. The top x documents returned by an initial query are identified as relevant. Various techniques are used to rank the terms and the top t terms of the top documents are added to the query. Parameters x , t , and the number of iterations can be tuned. Text REtrieval Conference (TREC) activities most commonly use this utility.
- *Semantic Networks* – This utility uses concept hierarchies where related concepts are linked to each other. The links have strength of relationship associated. This hierarchical network is used to expand queries and documents, to contain more terms describing the contents of the query.
- *Clustering* – Initially, the documents are grouped into clusters to reduce search space. Afterwards, the query is run against clusters believed to be relevant. In the case of automatic clustering, the main drawback of this utility is that a clustering algorithm may have performance issues. For that reason, this utility is not widely used.
- *Thesauri* – This involves automatic or manual generation of thesauri, then thesauri is used to expand queries or documents in order to improve information retrieval.
- *Parsing* – Parsing rules (e.g., stemming, exclude common prefixes and suffixes, include common phrases as terms, etc.) are used to improve the quality of information retrieval. Parsing plays a critical role in all information retrieval systems.
- *N-grams* – The terms in a query are decomposed into n-grams (i.e., sequences of n characters); that are matched against documents. N-grams are used for detection and correction of spelling and optical character recognition (OCR) errors, text compression, and are language independent. Some improvement in an inverted index is needed with this utility.

Those described retrieval utilities, and others, can be integrated with any retrieval strategy, independently or in combination with other retrieval utilities. There is no clear answer regarding which utility, or combination of utilities, and strategy is the best. This also requires further testing and research.

2.2.3. Information Retrieval Efficiency

In addition to finding relevant documents for a query, it is very important to determine how long it takes to find the documents. Relying on a constant increase of machine speed is not good enough. Instead, there is a need for special techniques that can improve the performance of information retrieval. The most common technique is to create inverted indexes. Note, indexing text differs from indexing relational structured data. With text, we are searching for documents that contain (or do not contain) certain terms or have terms in close proximity of each other. We are not interested in the text as an atomic unit as is the case with relational data. Therefore, b-tree and bitmap indexes designed for a relational search cannot be used in a text search. An inverted index is used for a text search. Text is pre-processed for all unique significant terms. The list of those terms is called an index. For each term, the inverted index maintains a list of documents that contain the term, with the offset for the location of the term in the document. An illustration of an inverted index, with the associated list, is:

$$Term_i \rightarrow (\{doc_1, offset_1\}, \{doc_2, offset_2\}, \{doc_{15}, offset_3\}, \dots, \{doc_n, offset_m\})$$

The downside to generating efficient inverted indexes is the need for additional processing and storage. However, improved query performance significantly outweighs this downside. Special techniques, such as index compression and index pruning, are used to reduce the storage required for inverted indexes.

Information Retrieval is the first step in text mining. Its goal is to find documents that may be of use in solving a problem.

2.3. Information Extraction

Information Extraction (IE) is a technology based on analyzing natural language with the aim to automatically extract structured information (i.e., contextually and semantically well-defined data from a certain domain) from unstructured machine-readable text documents. Figure 3 gives an example of Information Extraction. IE typically accepts text as input and, using rule modules and/or a statistical model, produces fixed-format, definite data as output that is stored in a spreadsheet or a database table for later analysis. Once this process is complete, the traditional data mining techniques can be applied to the structured output. Output can also be used for indexing purposes within such Information Retrieval applications as Internet search engines Yahoo and Google. IE alone can serve as an effective approach to text mining if the knowledge to be discovered is expressed directly into the documents to be mined.

The main difficulty in Information Extraction comes from linguistic complexity. That is, there are many ways to express the same fact.

This is an example of how natural language can express the same reality in different ways:

- Bank ABC has raised its lending rate by half-a-percent.
- Based on a half-a-percent lending rate increase by their competitors, bank ABC has followed along.
- Bank ABC will now start charging 4.25% on its lending rate from 3.75%.
- Petar Petrovic, Chief Financial Officer of bank ABC, announced the new half-a-percent lending rate increase to the media today.

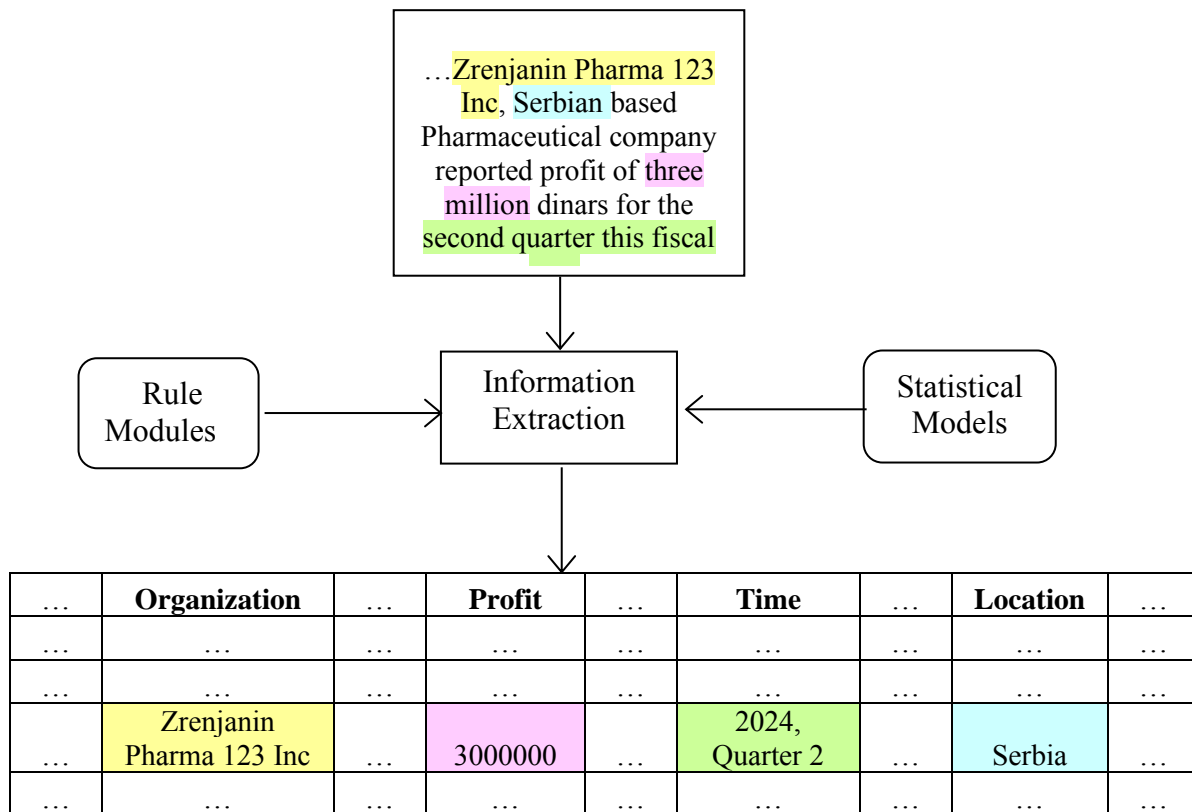


Figure 3. An Example of Information Extraction Process

Element Types in Information Extraction. There are four main types of elements that can be extracted from text:

1. *Entities* – The basic building blocks that can be found in documents. Some examples of entities are person, organization, location, money, gene, and drug.
2. *Attributes* – Attributes are characteristics of the extracted entities. Examples include the age of a person, the type of drug, the address of a location, and currency of the money.
3. *Facts* - The relations that exist between entities. Examples include a family relationship between two people, or a diagnostic relationship between symptoms and illnesses.
4. *Events* – The activities or occurrences of interest in which entities participate. Examples include anniversaries, mergers between two organizations, or a new drug invention.

2.3.1. Information Extraction Tasks

The most comprehensive work related to codifying and expanding IE tasks has come from Message Understanding Conferences (MUC). MUC identified five main IE tasks [7] [10]:

1. *Named Entity Recognition (NE)* – NE finds entities (e.g., the names of people, locations, medications). It is a very basic and most reliable IE phase. The accuracy of the extraction results on the NE task can be up to 95-percent, which is close to a human level. This task is weakly domain dependent (i.e., changing the subject area or domain of the text may or may not degrade the performance levels).
2. *Template Element (TE)* - TE adds descriptive information to NE results – that is, attaches attributes to the entities. The accuracy of TE is around 80-percent, so there is room for improvement, since human accuracy is about 90-percent. Similar to NE, a TE task is weakly domain dependent.
3. *Template Relationship (TR)* – TR finds relationships between entities. To find those relationships, it doesn't look in the text; it is using the template elements extracted from text during TE task phase. The accuracy of TR is up to 75-percent. TR task is also weakly domain dependent.

4. *Scenario Template (ST)* - ST finds events that the entities participate in, and tie TE and TR results into specified event scenarios. Scenario templates are prototype outputs of IE systems. The accuracy of ST is quite low (i.e., around 60-percent) because it is a difficult task. Human accuracy for this task is approximately 80-percent, which illustrates the intricacy involved. These figures should be taken into account when considering appropriate applications of ST technology.

5. *Coreference Task (CO)* – CO identifies which entities and references (e.g., pronouns) refer to the same thing. The importance of this task is the way it is a building block for TE and ST. CO resolution is not a precise process, particularly when applied to the solution of anaphoric reference. CO results vary broadly and most commonly achieve accuracy rates of 50-60-percent. TE task is also weakly domain dependent.

Some of the most prominent techniques that are applied to the tasks in information extraction involve probabilistic models: hidden Markov models (HMM), stochastic context-free grammars (SCFG), and maximal entropy (ME).

2.3.2. Information Extraction Architecture

Feldman and Sanger identified four foremost components in a typical IE architecture: tokenization, morphological and lexical analysis, syntactic analysis, and domain analysis, as shown in Figure 4 [7].

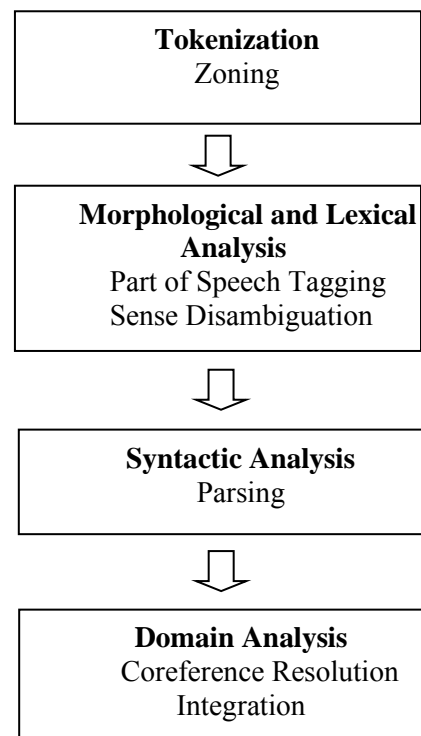


Figure 4: Architecture of a typical Information Extraction system.

Such components can also be seen as main phases in IE. First, tokenization divides the text into simple building blocks (i.e., tokens), such as numbers, words, sentences, paragraphs, sections and chapters. The second component, morphological and lexical analysis, include activities such as dropping common words (e.g., the, an, that), stemming, lemmatization, tagging, creating basic phrases, and clarifying ambiguous words and phrases.

Stemming is a heuristic process that cuts off the prefixes and suffixes of words with the intention to reach their roots. For instance, the following words: learning, learner, pre-learn, and learn are stemmed to the root word: learn.

Lemmatization is a process of determining the lemma or standard form for a given word, with the use of a vocabulary and morphological analysis of words. The word *good* is the lemma for words *best*,

good, and *goodies*. Part of speech (POS) tagging labels every word with a part of speech tags (e.g., NN - noun, VB - verb, VBD verb past tense, and PP personal pronoun). After identifying basic lexical elements in the text, the first step in the next component is syntactic analysis, which identifies entity types such as names, locations, organizations, and times. The words in the sentence are scanned while trying to match one of the patterns in the predefined set of regular expressions. Shallow parsing identifies the noun and verb groups, while deep parsing specifies their internal structure and role within the main sentence. Next, relations between entities are extracted by using domain specific patterns. Relations are usually considered at a sentence level. The patterns are developed using linguistic analysis and domain expert knowledge. Coreference, or anaphora resolution, is an important part of relationship extraction. It is a process of checking whether NLP expressions refer to the same entity. Anaphora resolution has two main approaches. The first approach is knowledge-based with manually encoded rules based on linguistic analysis. It is a very tedious and time consuming and, generally, doesn't result in a robust system. The second approach, which is more prominent and the most successful, is machine learning based on a human annotated set of documents. Refer to Table 2 for the main characteristics of both approaches.

Table 2:
Knowledge Engineering versus Learning Systems

Knowledge Engineering	Learning Systems
Rule based.	Use statistics or other machine learning.
Developed by experienced language engineers.	Developed by programmers who do not need language engineering experience.
Require small amount of training data.	Require large amounts of annotated training data.
Development can be very time consuming.	Development takes considerably less time.
Some changes may be hard to accommodate.	Some changes may require re-annotation of the entire training corpus.
Make use of human intuition.	

Advanced IE Systems. Other advanced IE systems use visual elements of a document, such as pictures, or whether the information is in the heading or footer. Such IE systems are called structural IE and are used to complement conventional IE processes.

Probabilistic models. The most common are hidden Markov models (HMM), stochastic context-free grammar (SCFG), and maximal entropy (ME).

2.4. Information Retrieval versus Information Extraction

The main differences between Information Retrieval (IR) and Information Extraction (IE) systems are [11]:

- IR systems find relevant texts and present complete text documents to the user.
- IE systems analyze text and presents only specific information that is relevant to the user.
- IE systems are more difficult and knowledge-intensive to build, being tied to particular domains and scenarios.
- IE is more computationally intensive than IR.
- In applications dealing with large text volumes, IE is potentially more efficient than IR because of the possibility of drastically reducing the amount of time that people spend reading texts.

- Consider the user of an IR system who wants information on bank mergers by entering a list of relevant words and receives a set of documents which contain matches. The user would then have to read the documents and extract the required information manually. Conversely, an IE system would automatically populate a spreadsheet with the names of relevant banks and their mergers.

2.5. Text Categorization

Text categorization refers to the assignment of natural language texts to one or more predefined categories based on their content [15]. Text categorization can be single-label, binary, and multi-label.

In single-label categorization, every document belongs to exactly one category. Binary categorization is a special case of a single-label category where the number of categories is two. In multi-label categorization, each document can belong to many categories. An illustration of document categorization is shown in Figure 5.

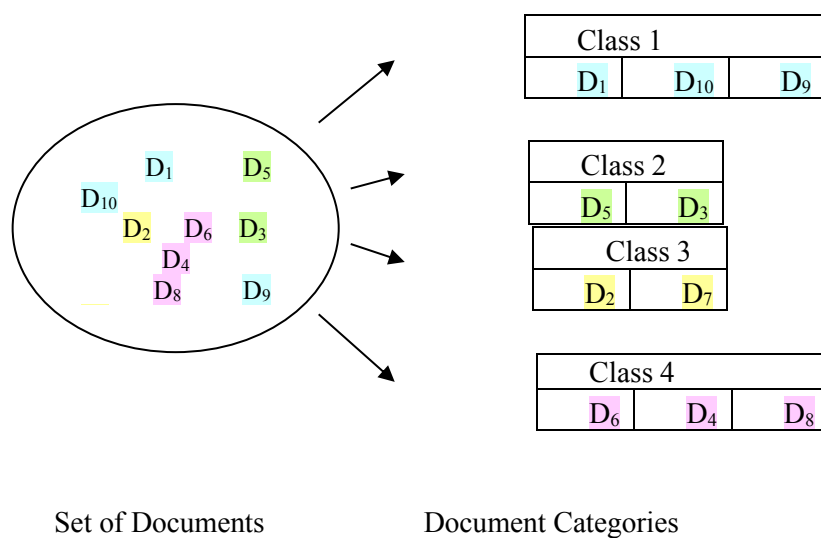


Figure 5. An example of single-label Text Categorization

Similar to the information extraction and other artificial intelligence tasks, there are two main approaches to text categorization: *knowledge engineering*, and *machine learning*. For knowledge engineering, the trained subject matter experts encode knowledge about categories manually into the system, either declaratively or (most commonly) in the form of classification rules. This approach can be quite expensive because it requires a lot of time and qualified people.

Machine learning is an automatic approach that builds a set of categories from pre-classified documents that are called training documents. There are four main issues to be considered with the machine learning approach: a decision on the categories that will be used to classify the instances, obtaining a training set (i.e., approximately 30 examples for each category), a decision on the features that represent each of the instances, and an algorithm for the categorization [7]. Types of algorithms include: *Probabilistic Classifiers*, *Bayesian Logistic Regression*, *Decision Tree Classifiers*, *Neural Networks*, *Regression Methods*, *Symbolic Rule Learning*, and *Support Vector Machines*.

In general, systems using the knowledge engineering approach provide better quality results. However, due to knowledge acquisition bottleneck, the machine learning approach is becoming more prominent for applications and research.

2.6. Text Clustering

Text Clustering is a technique that assigns each of the documents from a collection into one, or more, smaller groups called clusters. Grouping of documents is based on examination of the words without any previous information. Good clustering occurs when every document is similar to a document from its cluster, and less similar to the documents outside of its cluster(s). Clustering is used in data analysis for document retrieval, image segmentation, and pattern classification.

There are two main types of clustering: *flat* and *hierarchical*. *Flat* clustering partitions documents into a flat structure in which each document is placed in only one set. *Hierarchical* clustering groups documents into a nested chain of partitions. Each type can be either *hard* or *soft*. In hard clustering, each document can belong to only one cluster; whereas in soft clustering, documents belong to clusters with a fractional score between 0 and 1. An illustrative example of flat hard text clustering is shown in Figure 6.

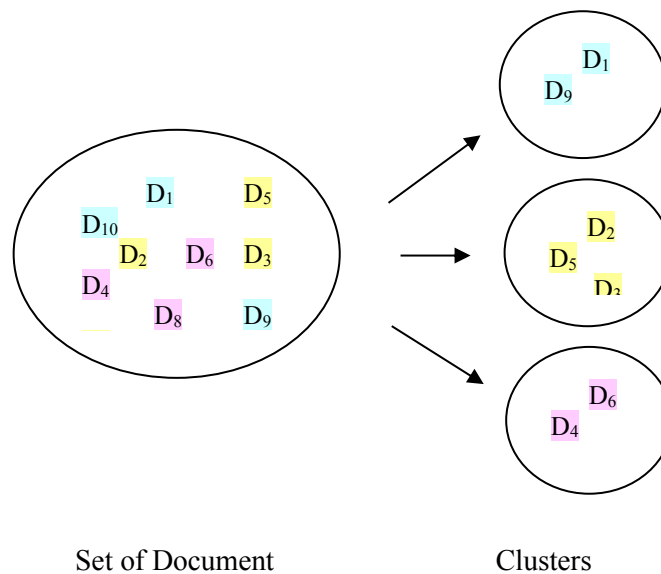


Figure 6. An example of flat hard Text Clustering

The documents need to be converted into vectors, in the feature space. The most common method is to use a bag-of-words document representation, where each word is a dimension inside the feature space. Each vector that represents a document has a component for each word. If a word is present in the document, then the component has a value of 1, or some other value; otherwise, 0. With a large document collection, the dimension feature space may be tens, or hundreds, of thousands. For that reason, feature selection is very important in terms of improving performance. The goal is to eliminate components (i.e., terms) that are irrelevant either due to infrequent occurrence or occur in a constant distribution among the different documents. *Latent Semantic Indexing* is based on applying Singular Value Decomposition (SVD) and is an increasingly popular technique used for dimension reduction.

The most commonly used algorithms for clustering are *K-means*, *Expectation Maximization* (EM), *Probabilistic Clustering*, *Hierarchical Agglomerative Clustering* (HAC), *Nearest Neighbour Clustering*, and the *Buckshot*. Consider the following details of a very prominent hard K-means algorithm.

K-Means Algorithm. For internal representation of documents, the vector-space model is used and the cluster is represented as the centroid of the documents belonging to that cluster. Initially, documents are grouped into k clusters, randomly, or chosen by an external heuristic. Hence, the algorithm proceeds, as shown in Figure 5 [11].

```

initialize cluster centroids to arbitrary vectors
while further improvement is possible do
  for each document  $d$  do
    find the cluster  $c$  whose centroid is most similar to  $d$ 
    assign  $d$  to this cluster  $c$ 
  end for
  for each cluster  $c$  do
    re-compute the centroid of cluster  $c$  based on documents assigned to it
  end for
end while

```

Figure 5. The K-Means Algorithm.

Formula for centroid M_i of a cluster C_i :

$$M_i = |C_i|^{-1} \sum_{x \in C_i} x$$

The K-means algorithm maximizes the clustering quality function Q :

$$Q(C_1, C_2, \dots, C_k) = \sum_{C_i} \sum_{x \in C_i} \text{Sim}(x - M_i)$$

This algorithm divides a collection of vectors into a set of clusters.

Since K-means is efficient and simple, it is, therefore, commonly used.

Text Clustering Evaluation. Comparing internal measures of clustering produced by different algorithms demonstrates which algorithm results in a better estimation of the general optimization problem. Other more relevant measures see how good clustering is for human use or for further processing. For this, the most common measure is *purity*.

$$Purity(C_i) = \max_j |L_j \cap C_i| / |C_i|$$

Where $\{L_1, L_2, \dots, L_n\}$ are the manually labelled classes of documents, and $\{C_1, C_2, \dots, C_m\}$ are the clusters returned by the clustering process. Other measures are *entropy* and *mutual information*. Yet, the most useful is the measure of the utility of the resulting clustering in its intended application [7].

2.7. Text Categorization versus Text Clustering

The main differences between text categorization and text clustering are:

- Categorization is a supervised process while clustering is an unsupervised process
- Categorization labels each document with a category, but clustering does not
- Categorization requires a rule that will accurately assign labels to documents, or training a classifier, while clustering groups documents into clusters based on how similar each is to another
- With categorization, we know how many classes there will be
- Although, with categorization, we assume that there will be a natural dissection, we don't know how many clusters there will be.

2.8. Core Mining Operations

Pattern Discovery. Four of the most common types of patterns found in text mining are *distributions*, *frequent sets*, *near frequent sets*, and *associations*. Distributions are looking for concept selection (i.e., a sub-collection of a complete document collection) and its behavior. A frequent concept set is a group of concepts represented in the document collection with co-occurrences at or above, a minimal support level. Near frequent concept sets find relations between frequent sets of concepts. Associations are

directed relations between concepts or set of concepts. Wong, Whitney, and Thomas define an association as an implication of the form $X \rightarrow Y$ where X is a set of antecedent items and Y is the consequent item [12]. An example of an association in data and text mining is that 90-percent of newcomers who open their first saving account with a bank will also obtain a credit account with the same bank.

Trend Analysis. Trend analysis in text mining refers to the analysis of concept distribution across several document subsets over time. This core mining operation is based on the notion that concept distributions are functions of document collections. For example, one can compare news articles about various banks' performances of the last and second last quarter of the current year. Such a comparison will highlight points where proportion changed between two time periods.

Ephemeral Association. Ephemeral associations are investigations on the influence of the most occurring concepts in a period on other concepts. Examples include news stories about sub-prime mortgages that correlate with the emergence of stories about new rules on risk reporting.

2.9. Data Mining versus Text Mining

The difference between regular data mining and text mining is that, in text mining, the patterns are extracted from natural language text rather than from structured databases [1]. Both text and data mining deal with a large volume of data and have the same ultimate goal to discover new trends and patterns for, most commonly, better decision making. Data mining, also known as Knowledge Discovery in Databases, is a more mature discipline that has lent a variety of its own techniques, methods, and algorithms to text mining. The main difference is that text mining deals with unstructured data, which introduces another level of complexity.

Metaphor 'mining' (in both data and text mining) is used to highlight the extraction of valuable knowledge from a large amount of data or information. Hearst argues that mining might not be the best metaphor and emphasizes that data mining applications do not discover new factoids; instead, these tend to be (semi) automated discoveries of trends and patterns across very large datasets, usually for the purposes of decision making [1].

2.10. Text Mining Applications

Text mining tools have been proven to work particularly well in the following three areas: corporate finance (e.g., business intelligence, noting trends, identifying correlations, researching references to specific transactions, corporate entities or persons, etc.), patent research (e.g., investigating patent development strategies, finding ways to exploit existing corporate patent assets even better), and life science (e.g., explore vast collections of biomedical research reports to identify complex patterns of interactivities between proteins). Other common text mining applications include survey analysis, crime pattern analysis, and call center data analysis. There are many Text Mining tools available, with some of the most prominent being: SAS (Statistical Analysis System) Text Miner, GATE (General Architecture for Text Engineering), IBM SPSS (Statistical Package for the Social Sciences) Modeler Text Analytics, IBM Watson, Apache OpenNLP, UIMA (Unstructured Information Management Architecture), Google Cloud NLP, Amazon Comprehend, Discover Text, Datavid Rover, R, Python, Oracle Miner, and DB2.

3. Contribution

This work is primarily an overview of text mining and its main techniques, including: information retrieval, information extraction, text classification, text clustering, and core mining operations. Future work into text mining can introduce tools to be evaluated with an expanded set of evaluation criteria. Tools to consider are SAS Text Miner, Datavid Rover, and Apache OpenNLP

4. Conclusions

The rapid development of computer systems has made it easy to store large amounts of data. The increasing number of text documents that exist on corporate networks, the Internet, FTP resources, news, email, social media and other online resources can be overwhelming. Traditionally, information systems deal with structured data. Text mining delivers an impressive attempt to understand natural language text. This is not a one-step process; rather, it involves many processes and such techniques as information retrieval, natural language processing, machine learning, information extraction, text categorization, text clustering, pattern identification, and trend analysis.

The most productive subject areas for text mining include corporate finance, patent research, life science, survey analysis, crime pattern analysis, and call center data analysis.

Text mining tools are crucial for dealing with large amounts of text. The main limitations with text mining models and techniques are due to the logical structure of text and the nature of the complex semantic relationship. Even though some of the tools appear to be very powerful, there are none that actually understand text. However, the rapid development of text mining tools along with other areas in data science continues to assure us that some of the limitations hampering these tools will be addressed in a relatively short period of time.

Acknowledgment:

I would also like to thank Professor Ivana Berkovic and the committee for reviewing this paper and for their valuable feedback. Also, for the opportunity to present this paper at the conference.

References:

- [1] Hearst M. "What Is Text Mining?". SIMS, University of California, 2003. Retrieved September 7, 2024 from URL: <http://people.ischool.berkeley.edu/~hearst/text-mining.html>.
- [2] Tun N.N. "Semantic Enrichment in Ontologies for Analysis and Matching". PhD Thesis. Japan Advanced Institute of Science and Technology, 2007. Retrieved on September 7, 2024 from URL: <http://www.jaist.ac.jp/library/thesis/is-doctor-2007/paper/nitun/paper.pdf>.
- [3] OpenKM Knowledge Center. "Understanding taxonomy, keywords, categories and metadata groups". Retrieved on September 12, 2024 URL: <https://docs.openkm.com/kcenter/view/okm-6.3-com/essential-fundamentals.html>.
- [4] Kimball, R. "The Data Warehouse Toolkit". John Wiley and Sons. New York, 1996.
- [5] Inmon, W. H. "What is a Data Warehouse?". Prism. Brighton. Vol.1, No. 1, 1995.
- [6] Sullivan, D. "Document Warehousing and Text Mining". Wiley Computer Publishing. John Wiley & Sons, Inc. New York, 2001.
- [7] Feldman R., Sanger J. "The text mining handbook". Cambridge University Press. New York, 2007.
- [8] Weiss S. M., Indurkha N., Zhang T. Text Mining. Springer. New York, 2005.
- [9] Grossman D.A., Frieder O. "Information Retrieval". Springer. Dordrecht, 2004.
- [10] Cunningham H. Information Extraction, Automatic. University of Sheffield, 2004. Retrieved September 7, 2024 from URL: <http://gate.ac.uk/sale/ell2/ie/main.pdf>
- [11] Chakrabarti S. Mining the Web. Morgan Kaufmann Publishers, 2003.
- [12] Wong P.C., Whitney P., Thomas, J.. "Visualizing Association Rules for Text Mining". Pacific Northwest National Laboratory, 1999. Retrieved on September 8, 2024 from URL: <https://ieeexplore.ieee.org/abstract/document/801866/metrics#metrics>
- [13] GeeksforGeeks, "Text Mining in Data Mining" 2024. Retrieved on September 10, 2024 from URL: <https://www.geeksforgeeks.org/text-mining-in-data-mining/>
- [14] Dumais S, Platt J., Heckerman D., Sahami M. "Inductive Learning Algorithms and Representations for Text Categorization". Microsoft Research, 1998. Retrieved on September 10, 2024 from URL: <https://www.microsoft.com/en-us/research/uploads/prod/1998/01/Inductive-Learning-Algorithms-and-Representations-for-Text-Categorization.pdf>

IT equipment and software for training, modeling and data analysis for flood and forest fire prevention, protection and management in project SOLVE

Andrijana Bocevska¹, Igor Nedelkovski¹, Aleksandar Markoski¹, Zoran Kotevski¹, Kostandina Veljanovska¹, Blagoj Ristevski¹ and Snezana Savoska¹

¹ University St Kliment Ohridski, IMaj nn, 7000 Bitola, R. North Macedonia

andrijana.bocevska@uklo.edu.mk; igor.nedelkovski@uklo.edu.mk; -
aleksandar.markoski@uklo.edu.mk; zoran.kotevski@uklo.edu.mk;
kostandina.veljanovska@uklo.edu.mk; blagoj.ristevski@uklo.edu.mk; snezana.savoska@uklo.edu.mk

Abstract:

Recent disasters around the globe strongly indicate that most challenging territories for managing floods and forest fires are the cross border ones. Various reasons (economic, social, political, cultural) make it difficult to agree, establish and implement joint strategies and policies dealing with climate change impacts, societies' resilience and emergency management. A huge identified deficit in exploitation of research and projects' outputs to strengthen civil protection systems exists at border areas. Available technological advancements and innovations do not reach fire brigade, forestry and civil protection services. This is due to lack of personnel, high rigidity in existing plans and procedures, inability to create cross border standard operation procedures and most importantly to identify the cross border area as a single area of intervention prior and during an emergency. Project Cross Border Complex Floods and Forest Fires Prevention and Management (SOLVE) focuses on joint actions for most common risks (forest fires and floods). This paper gives an overview of the modern IT equipment and software for training, modeling and data analysis for flood and forest fire prevention, protection and management acquired within this project. The paper also, emphasizes future potentials of the equipment in terms of using data gathered in the project for prediction in combination with potentials of machine learning.

Keywords: floods, forest fires, cross border, IT equipment, machine learning in fire/flood prediction

1. Introduction

Recent global disasters underscore the complexities faced in managing floods and forest fires, particularly in border regions. Economic, social, political, and cultural differences pose barriers in establishing joint strategies and policies for climate change adaptation, societal resilience, and emergency management. These challenges have resulted in an identified deficit in leveraging research and project outputs to fortify civil protection systems, hindering the adoption of technological advancements by fire brigades, forestry, and civil protection services.

The Republic of North Macedonia and Greece came together to strengthen their civil protection services and emergency management by initiating the "Cross Border Complex Floods and Forest Fires Prevention and Management" project with the acronym SOLVE. This project is an indication how borders can benefit rather than limit neighboring nations, especially in times of crisis and emergency.

The project's objective is to minimize risks from forest fires and floods in the long term, transforming cross-border regions from vulnerability to resilience. By implementing joint actions, conducting large-scale exercises, and emphasizing immediate results utilization, SOLVE sets a good practice example for effective disaster risk reduction that emphasizes cross-border cooperation.

Main outputs of the project were: a) Modern equipment and joint training of operational teams (1st responders), b) Provision of scientific support and knowledge (through innovative approaches,

methodologies and tools to operational decision making, c) Capitalization and mainstreaming of results, and d) active citizens' participation to prevention of activities through Participatory GIS platform and focused actions. Such approach is novel and fosters the cross border cooperation at a sustainable basis, following the provisions of new civil protection law 4662/2020 and the 2020 cross border collaboration guidelines of DG ECHO.

SOLVE has a dual character approach: a) Working on most common risks, building on past and ongoing successful projects' results and b) joint actions not only at the levels of risk assessment, early warning and management procedures, but also at conducting large scale joint exercises, common training and immediate exploitation of results both at policy and operational levels. The overall SOLVE objective is to minimize the risks from forest fires (including Wild Urban Interface ones) and Floods (including flash and urban floods) for the Florina regional unit & Pelagonia region in a long-term basis and in a way that can be replicated in other cross border areas.

The project's lead partners were the Decentralised Administration of Epirus and West Macedonia based in Ioannina, Greece, the Center for Development of Pelagonia Region, University St. Kliment Ohridski – Bitola (UKLO), and the University of Western Macedonia.

About 45 professors, involved from UKLO, shared their academic expertise and scientific knowledge with the community and relevant institutions, with the aim of increasing risks and preparedness for attacks by forest fires and floods at five levels. The first level included fire and flood prevention and was based on risk assessment. The second one was protection and included the application of structural and eco-system measures to reduce the dangers of forest fires and floods. The third level was preparedness and included the activities of hazard forecasting, early warning and preparation of appropriate emergency plans. The fourth level was response and included activities related to how to raise the quality of stocks immediately after the occurrence of a forest fire or flood. The last level titled recovery was concerned with identifying measures for short and long term recovery of areas affected by fires and floods.

Faculty of Information and Communication technologies (FICT) as a unit of UKLO was responsible for the preparation of the technical specifications and technical offer of modern IT equipment and software for training, modeling and data analysis for prevention, protection and dealing with floods and forest fires. These deliverables were carried out in 4 lots in Bitola, with financial assistance from Interreg IPA Cross-Border Cooperation Program "Greece - Republic of North Macedonia 2014-2020".

The rest of the paper is structured as follows. The next main section of the paper describes the technical specifications and technical offer about IT equipment and software within the SOLVE project. The content is subdivided and presented according to the 4 project lots in Section 2.1, Section 2.2, Section 2.3 and Section 2.4, respectively. We conclude in Section 3 and discuss potentials of the equipment in combinations with prospects of machine learning in Section 4.

2. Technical specifications and technical offer about IT equipment and software

The management of disasters and emergencies is a multi-faceted task that requires involvement of a wide range of actors. These actors typically include public authorities, civil protection authorities, local governments, police, fire brigade, first responders, volunteers, and critical infrastructure operators. The entirety of aforementioned actors shall be equipped with a mixture of competencies, in order to prove effective during crises. The basis of the project implementation is the exploitation of the emerging IT equipment and software for training, modeling and data analysis for prevention, protection and dealing with floods and forest fires. For the purpose of the project the preparation of the technical specifications and technical offer about IT equipment and software was made in 4 lots i.e.:

- **Virtual reality mobile training center** (Projection Dome, Fire simulator training software, Laptops),
- **Software tools for citizen participation in civil protection**
- **Specialized equipment for analysing flooding data** (aerial vehicle – drone, wireless sensor network with LoraWAN sensors for water level detection and flood detection)
- **Specialized equipment for analysing forest fire data** (aerial vehicle – drone, wireless sensor network with LoraWAN sensors for forest fire detection)

2.1. Virtual reality mobile training center – Lot No.1

The basis of the Lot No.1 is the exploitation of the emerging technology of Virtual Reality. This technology has grown rapidly in recent years and the ability to fully immerse into virtual worlds and realistic experiences makes it highly effective in educational, experiential or training applications. According to [1], people remember 10% of what they read, 20% of what they hear, 30% of what they see, and 90% of what they do. Therefore, VR is an incredibly powerful tool for training so far as it is based on a learning-by-doing approach. For that purpose, a **3 meter open faced dome** was procured within the project [2]. For collaborative and immersive learning experiences to students the dome is placed in the Center for Virtual Engineering in the Faculty of information and communication technologies - Bitola, Figure 1. Its dimensions are the following ones: height: 3 m, area 7.1 m², screen 14,1 m², capacity (seating/standing) 5/7.

The dome theater is a new format of seeing and showing things, where a VR immersive experience is not individualized, but actually, a shared social activity. When it comes to digital dome projection, the quality of the projected image is of the utmost importance.

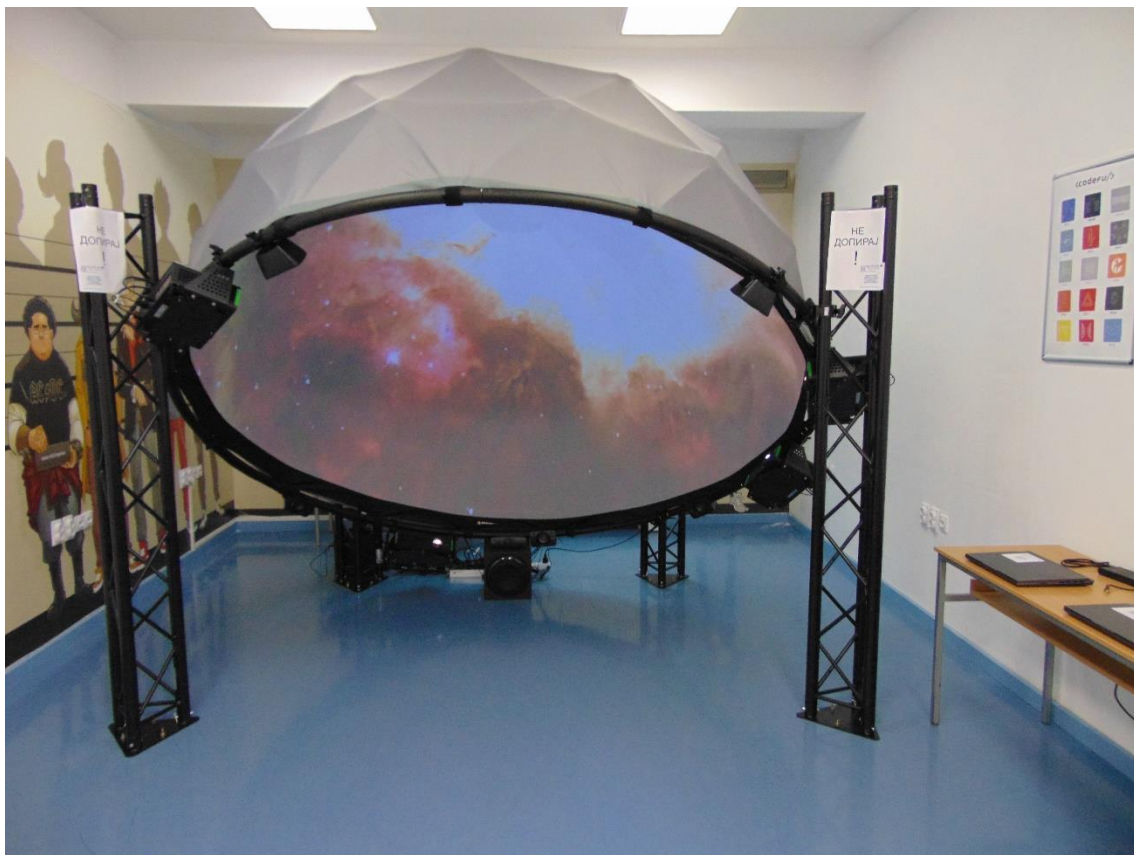


Figure 1: 3 meter open faced dome

There are many factors that influence the final image, but three components of the dome projection system which can be particularly emphasized are: the screen, the projectors, and the media server.

- **The screen** - is the thing the whole audience is looking at, but ideally, they should not really see the screen, just the image. Negative-pressure technology, suspends the screen evenly inside the dome, stretching it outwards to create a wrinkle free surface. Screen gain, the amount of light reflected off the surface, is critical. In a dome, the reflected light falls on other parts of the screen, lowering the contrast of the image, whereas in a regular cinema the reflection would hit the walls, ceiling, and floor. Therefore, a dome screen is not white, as might be expected, but grey. Exactly what shade of grey depends on the size of the dome and the projectors. Because projectors are placed around the perimeter of the dome in proximity to the screen, all the fabrics are tested and certified as a flame retardant.

- **The projectors** - the system consists of four digital light processing (DLP) projection modules mounted around the dome perimeter on a rig with resolution 3K. FP HD X4 provides an image quality that is six times better than that of a regular single-channel projection system. A fully immersive effect and an unforgettable experience are guaranteed with these projectors.
- **The media server** - the heart of the projection system is the media server (this dome uses a single-server solution). Unlike some systems where there is one media server for each projector, with content pre-split for each, another server to synchronize everything, and perhaps another for calibration and blending, this single-server solution supports up to 64 projectors. This includes fully automatic camera-based calibration, which eliminates the need for manual masking, blending and pre-splitting content. Fulldome.pro media server supports Fulldome live real-time content input from external sources.

After installing the projectors and pointing them in roughly the right direction, a camera with a fisheye lens is placed in the center of the dome. The operator connects the camera to the single media server and starts the calibration process. Fully-automated calibration software takes care of the rest. Each projector displays a series of patterns on the screen, which are captured by the camera and analyzed. The software works out where each projector overlaps with others and maps the intensity over the entire screen and a full calibration is completed in under 15 minutes. In most cases, calibration only needs to be performed once for a perfect dome image. But there are times where a few tweaks may be required. Settings can be adjusted to compensate for stray light, objects inside the projection dome, and other factors. Once these parameters are set, the system will be able to calibrate despite the interference.

An extensive fulldome content starter package is included with the purchase of every Fulldome.pro projection system.

An integral part of Lot No.1 represents a **software for Fire Simulation and Training SimUshare** [3]. SimUShare is a state-of-the-art fire simulator software package, designed to produce ultra-realistic simulations including fire incidents, hazardous material situations, and a wide range of other emergency scenarios. With incredible realism, it delivers a truly immersive training experience, improving the decision-making skills that are vital for emergency responders. Screenshot of creating a basic simulation using this software is presented in Figure 2.



Figure 2: Creating a Basic Simulation with SimUshare

The features that make this software the most versatile fire simulator on the market are [4]:

- **Possibility to use own pictures** – a realistic situations can be created by including pictures of buildings and structures within response area. Also images of apparatus and tools for a fully immersive experience can be added.
- **Support of critical training requirements** - SimUshare's customizable training environment

creates the conditions and elements needed for company officer, incident command, and size-up training.

- **Possibility to conduct remote or in-person real-time simulation training** - flexibility is key and SimUshare allows to conduct real-time simulation training, regardless of participant location – virtual or on-site.
- **Possibility to embed the simulations** – simulations can be embedded into learning management systems, websites, or emails for easy access.
- **Possibility to build simulations across all industries** – first response training goes beyond fire departments. SimUshare is the proven leader in other first-response environments like the military, industrial safety, schools, and more.
- **Interactive walkthroughs** – interactive walkthroughs of any building through a simple click of a button can be created.
- **Promotional exams** – SimUshare is the perfect tool to create promotional exam scenarios that provide instant performance feedback to the team members.
- **Simple to use** - Unlike others, SimUshare is incredibly intuitive and easy to learn. There is nothing to download and they offer helpful tutorials. Highly responsive support team is also at disposal.
- **Evolve simulation conditions** – SimUshare allows to configure simulations to change conditions on the ground based on elapsed time or participant decisions.
- **Built-in checklist functionality** – one can easily create helpful checklists, so participants can follow prescribed SOP's and protocol.
- **Internal messaging** – SimUshare's built-in messaging function allows participants to communicate with each other in real-time.

For training, modeling and data analysis for flood and forest fire prevention, protection and management in project SOLVE seven laptops Legion Slim 5 16APH8 were also procured [5]. Their configuration is as follows: Processor: AMD Ryzen 7 7840HS, Graphics: NVIDIA GeForce RTX 4070 8GB GDDR6, Max Memory: Up to 32GB DDR5-5600 offering, Storage Type: M.2 2280 SSD 1TB, Display: 16" WQXGA (2560x1600) 300nits, Camera: FHD 1080p, with E-shutter, fixed focus, Warranty: 24 months, Operating System: Installed licensed Windows 10 pro.

2.2. Software tools for citizen participation in civil protection– Lot No.2

Participatory **GIS application “SOLVE”** for participation of citizens in civil protection prevention, preparedness and management activities was developed for the Lot No.2, Figure 3.

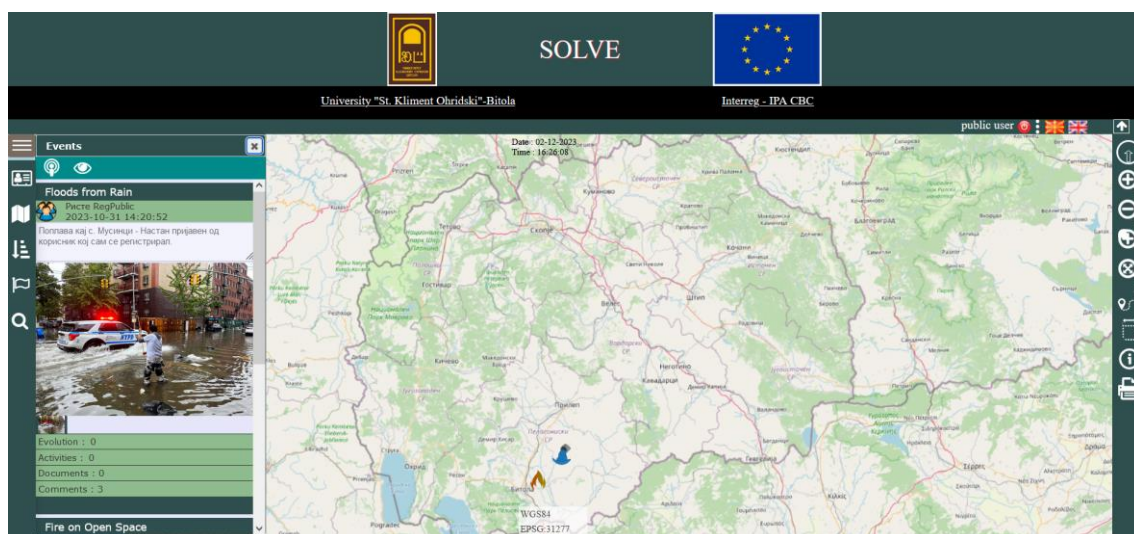


Figure 3: Main page of the application “SOLVE”

Authorities who have access can publish events with all specific details that will be publicly visible for all users. Users (registered or not registered) can put comments with pictures for the published events but will not be able to change and add official data of the event. Gathered information from them will be only through comments to avoid abuse and mislead. Particular users, groups or agencies can be given rights by super-administrator to change and add official data.

For each published event, a corresponding icon is displayed on the map at the location of that event. If multiple images are uploaded for a given event, each image can be viewed separately by clicking on the corresponding thumbnail in the bar below the large image window.

“SOLVE” is GIS application which contains geospatial maps and also allows super-administrators to introduce new geospatial maps through settings menu. Every information for events is time tagged and date and time of publishing is clearly shown so development of each event can be followed.

“SOLVE” is developed as WEB, Android and IOS application hosted on solve.mk domain [6].

2.3. Specialized equipment for analysing floods– Lot No.3

Specialized equipment for analysing floods (Lot No.3) consist of one aerial vehicle – drone (DJI Mavic 3), four pices of Milesight EM500-SWL LoRaWAN and two of Ultrasonic Distance/Level Sensor EM500-UDL sensors for water level detection and flood detection.

The **DJI Mavic 3 drone** can be a valuable tool in managing and responding to flood situations, Figure 4. It stands as a testament to the drone industry’s phenomenal growth and innovation. This high-end consumer drone is a marvel of engineering and design, showcasing a perfect balance between advanced features and user-friendly functionality. One of the most striking features is its dual-camera setup, offering both a wide-angle and a telephoto lens, ensuring that aerial photographers and videographers have more creative options than ever before. With a robust flight time, the Mavic 3 ensures remaining in the air longer, capturing those critical shots.

By using a comprehensive set of sensors, the Mavic 3 can navigate complex environments with ease, making it safer for both beginners and professionals. The O3+ Transmission system guarantees a stronger, more stable connection between the drone and controller, even in environments filled with interference. Furthermore, its compact and foldable design ensures portability without compromising on the features.



Figure 4: DJI Mavic 3 drone

The DJI Mavic 3 has a maximum Flight Time of 46 minutes and maximum Hovering Time of 40 minutes in an environment with windless conditions [7].

In Flood Situations, DJI Mavic 3 can play a role in following scenarios:

- **Damage assessment:** The Mavic 3's ability to fly over flooded areas provides a clear view of the extent of flooding, which is crucial for assessing damage to infrastructure and properties.
- **Monitoring water levels:** By capturing images and videos over time, the drone can help monitor changes in water levels, which can be used to predict future flooding or understand the impact of ongoing flooding.
- **Search and rescue:** Similar to its role in fire situations, the Mavic 3 can aid in search and rescue operations by identifying people stranded in flooded areas or assessing conditions from a safe distance.
- **Infrastructure inspection:** The drone can inspect critical infrastructure such as bridges, roads, and dams to identify potential damage or weaknesses that could exacerbate flooding.
- **Data collection:** It helps gather valuable data for research and planning purposes, aiding in better preparation for future floods and improving response strategies.

EM500 series is a sensor mainly used for outdoor environment through wireless LoRa network. EM500 series sensors is made up of a LoRa transceiver and a sensor. Among them, ultrasonic sensors and gas sensors are combined with LoRa transceiver.

EM500 device is battery powered and designed for multiple mounting ways. It is equipped with NFC (Near Field Communication) and can easily be configured by a smartphone or a PC software.

Sensor data are transmitted in real-time using standard LoRaWAN protocol. LoRaWAN enables encrypted radio transmissions over long distance while consuming very little power. The user can obtain sensor data and view the trend of data change through Milesight IoT Cloud or through the user's own Network Server. Features include [8]:

- Up to 15 km communication range;
- Easy configuration via NFC;
- Standard LoRaWAN support;
- Milesight IoT Cloud compliant;
- Low power consumption with 19000mAh replaceable battery.

EM500 series can be monitored and configured via ToolBox App or ToolBox software. LoRaWAN settings are used for configuring the transmission parameters in LoRaWAN network. EM500 devices support configuration backup for easy and quick device configuration in bulk.

EM500 series can be managed by Milesight IoT Cloud platform. Milesight IoT cloud is a comprehensive platform that provides multiple services including device remote management and data visualization with the easiest operation procedures. Before operating following steps, a Milesight IoT Cloud account is needed.

Submersible water level sensor EM500-SWL, Figure 5, is designed for measuring liquid level in harsh environments and transmitting data using LoRaWAN technology.

With this low power consumption technology, EM500-SWL can work up to 10 years with 19000 mAh battery [9]. Combining with Milesight LoRaWAN gateway and Milesight IoT Cloud solution, users can manage all sensor data remotely and visually. EM500-SWL is widely used for applications like tank level monitoring, river level monitoring, etc.

Ultrasonic Distance/Level Sensor EM500-UDL, Figure 6, is designed to make measurements of various kinds of liquids or objects much easier through ultrasonic waves, making it applicable in industries of many kinds. Supported with ultrasonic waves, a non-contact detection technology, the device is able to detect a wide range of liquids, including corrosive chemicals while providing impressive measuring precision of up to $\pm 1\%$ accuracy [10]. This allows the device to integrate seamlessly with numerous industrial applications, especially in the plastic, petroleum, fertilizer, and other industries that involve large usage of chemicals. The device offers an extensive measurement range of 0.25 to 10m with a variety of probing options depending on the choice. It is also highly competent in measuring containers of distinct sizes, including containers with small openings. The device can be tailor-made for diverse scenarios. The device is boasted with an exceptional IP67 and UV resistance performance that allows it to perform superbly even in harsh working environments, making it ideal for industrial and outdoor applications like ski resorts, manufacturing industries, and more.



Figure 5: Submersible water level sensor EM500-SWL



Figure 6: Ultrasonic Distance/Level Sensor EM500-UDL

2.4. Specialized equipment for analysing fire – Lot No.4

Specialized equipment for analysing fire (Lot No.4) consists of one aerial vehicle – drone (DJI Mavic 3) and ten pieces of Carbone Dioxide Sensor (4 in 1) EM500-CO2.

DJI Mavic 3 in Fire Situations, can play a role in following scenarios:

- **Aerial surveillance:** The Mavic 3 can provide real-time aerial views of fire-affected areas. This helps firefighters assess the extent of the fire, identify hotspots, and determine safe routes for firefighting efforts.
- **Mapping and planning:** The drone's high-resolution camera and advanced imaging capabilities allow for the creation of detailed maps and 3D models of the fire-affected areas. This helps in planning evacuation routes, resource allocation, and strategy formulation.
- **Search and rescue:** Equipped with thermal imaging, the Mavic 3 can assist in locating people who might be trapped or in distress, especially in low visibility conditions or dense smoke.
- **Damage assessment:** After the fire, the drone can be used to assess damage and gather information for insurance claims, recovery planning, and rebuilding efforts.

EM500-CO2, Figure 7, is designed for monitoring CO2 concentration, temperature, humidity and barometric pressure in outdoor scenarios. EM500-CO2 is designed for measuring gaseous carbon dioxide (CO2) concentration in harsh environments. It is useful in applications where knowing CO2 level is important such as e.g., greenhouse, building ventilation, fruit and vegetable storage. Besides, it also supports temperature, humidity and barometric pressure measurement. EM500-CO2 can function properly even in demanding outdoor and harsh environments [11].



Figure 7: Carbone Dioxide Sensor (4 in 1) EM500-CO2

3. Conclusions

University St. Kliment Ohridski – Bitola, together with the Center for the Development of the Pelagonian Planning Region - Bitola and two partners from Greece, the University of Kozani and the Region of Western Macedonia, with the project SOLVE joined with the aim of increasing the capacities of the region to deal with forest fires and floods. SOLVE created a coherent risk prevention and

management strategy and capacity building framework to exploit on existing or underdeveloped prevention and mitigation strategies and tools. SOLVE approach transforms cross border area from problematic one concerning disaster risk reduction and management into a promising one fostering and facilitating cooperation. With this project, UKLO actively engaged with its knowledge to contribute to the community's efforts to reduce the risks of forest fires and floods. It received modern IT equipment and software for training, modeling and data analysis for flood and forest fire prevention, protection and management worth more than 100,000 euros. With this project, UKLO represents a regional leading institution in the areas of risk assessment and strengthening preparedness for dealing with natural disasters.

4. Further Research

As an addition to the work done in project SOLVE, we would like to discuss the prospects of the equipment procured in the project in correlation to the potentials of machine learning in leveraging the data gathered from the sensors and drones and also software application created in the project.

There are numerous scientific efforts where artificial intelligence and machine learning are used in the field of fire and flood management. Authors of [12] invented the concept of Intelligent Software Agents (ISA) for forest fire prevention and fighting. In [13] a decision support system for managing forest fire casualties is created. The system integrates GIS technologies and utilizes a common user interface based on semi-automatic satellite image processing (fuel maps), socio-economic risk modelling and probabilistic models that would serve as a useful tool for forest fire prevention, planning and management. Federal Emergency Management Agency (FEMA) uses Software for Flood Mapping whose data could be used to implement prediction models of flood prevention [14].

Machine learning algorithms have the potential to recognize patterns and correlations within data, so prediction of the possibility of fires and floods or the progression of their expansion could be done. Similarly, the vulnerability of particular regions to fire/flood could be estimated. Moreover, predictions can then be used to inform management and policy decisions, as well as assist in the planning of response actions [15].

Fire and flood risk management is only effective if all stakeholders contribute their resources under a common goal of increasing capacity and reducing vulnerability. Machine learning plays a significant role in predicting forest fires and floods by analyzing vast amounts of data and identifying patterns that can indicate the likelihood of these events.

Our sensors could be used to detect fire or flood, drones could be used for remote oversight, but in order to implement machine learning driven solution for fire and flood management there is no unique solution. Consequently, we have to explore the possibilities of our equipment, to enlarge the scope of data acquisition from different sources, e.g. weather forecast, satellite images, etc. and define the model of machine learning for fire and flood prediction. Further research in this field will open potentials for even earlier reaction on the emerging disaster and effective management, lowering the risks and performing faster and more successful regeneration.

References:

- [1] L. Janoska, What Really Is The Cone Of Experience?, eLearning Industry, 2017. URL: <https://elearningindustry.com/cone-of-experience-what-really-is>.
- [2] 360 PROJECTION DOMES, Fulldome.pro, URL: <https://fulldome.pro/domes/>.
- [3] SimUShare Industry Leading Fire Simulation and Training Software, URL: <https://SimUshare.com/>.
- [4] SimsUshare Features, SimUShare, URL: <https://SimUshare.com/features-2/>.
- [5] Lenovo Product Specifications Reference, URL: https://psref.lenovo.com/Product/Legion/Legion_Slim_5_16APH8.
- [6] GIS application "SOLVE", URL: <https://solve.mk/>.
- [7] DJI Mavic 3 Specifications, URL: <https://dofly.com.pk/dji-mavic-3-specs/>.
- [8] Outdoor Environment Monitoring Sensor Featuring LoRaWAN EM500 Series User Guide, URL: <https://resource.milesight-iot.com/milesight/document/em500-series-user-guide-en.pdf>.

- [9] Submersible Water Level Sensor Featuring LoRaWAN EM500-SWL Specifications, URL: <https://resource.milesight.com/milesight/iot/document/em500-swl-datasheet-en.pdf>.
- [10] Ultrasonic Distance/Level Sensor Specifications, URL: <https://www.milesight.com/iot/product/lorawan-sensor/em500-udl>.
- [11] LoRaWAN® Carbon Dioxide Sensor (4 in 1) Specifications, URL: <https://www.milesight.com/iot/product/lorawan-sensor/em500-co2>.
- [12] Jaber A., Guarnieri F., Wybo J. L., Intelligent software agents for forest fire prevention and fighting, *Safety Science*, 39 (2001) 3–17, [https://doi.org/10.1016/S0925-7535\(01\)00021-2](https://doi.org/10.1016/S0925-7535(01)00021-2).
- [13] Bonazountas M., Kallidromitou D., Kassomenos P., Passas N., A decision support system for managing forest fire casualties, *Journal of Environmental Management*, Science Direct, Volume 84, Issue 4, September 2007, pp. 412-418, URL: <https://doi.org/10.1016/j.jenvman.2006.06.016>.
- [14] The Federal Emergency Management Agency (FEMA), May 6, 2024, URL: <https://www.fema.gov/about>.
- [15] Giannakidou S., Radoglou-Grammatikis P., Lagkas T., Argyriou V., Goudos S., Markakis E. K., Sarigiannidis P., Leveraging the power of internet of things and artificial intelligence in forest fire prevention, detection, and restoration: A comprehensive survey, *Internet of Things*, Elsevier, Volume 26, July 2024, URL: <https://doi.org/10.1016/j.iot.2024.101171>.

Digital platform for monitoring and forecasting the environmental situation of the Baikal natural territory

Igor V. Bychkov¹, Evgeniy S. Fereferov¹ and Roman K. Fedorov¹

¹ *Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences, 134 Lermontov str, Irkutsk, 664033, Russia*

bychkov@icc.ru; fereferov@icc.ru; fedorov@icc.ru

Abstract:

The article presents the results of research carried out within the framework of a large scientific project supported by the Ministry of Science and Higher Education of the Russian Federation. The project is aimed at creating new methods and technologies for integrated environmental monitoring and forecasting based on digital platforms that collect, store, process, analyze large arrays of heterogeneous spatial and temporal data, as well as a set of mathematical and information models, services and machine learning methods and their validation for the Baikal Natural Territory. The authors have developed a digital platform that is the integrating basis of digital environmental monitoring. The digital platform provides accumulation, exchange and management of environmental monitoring data in a structured form. The digital platform also provides the development and execution of services for all participants of the digital ecosystem (monitoring participants).

Keywords:

Digital monitoring, forecasting, digital platforms, web-services, machine learning

1. Introduction

The Baikal Natural Territory (BNT) is a complex natural complex, where Lake Baikal, recognized by UNESCO as a World Natural Heritage Site, is located. There are a number of environmental problems in the BNT. Large cities and industrial centers located in the southern part of the BNT are sources of air, water and soil pollution, which in turn affects the health of the population. In addition to environmental problems of anthropogenic nature, there are dangerous geodynamic processes in the BNT (seismicity, landslides, rockslides, mudflows, etc.), as well as problems with the state of boreal forests (including epidemiological damage, spread of pests, natural fires, climate change, etc.). The development of means and methods of comprehensive geo-ecological control is an important task for the conservation of Lake Baikal ecosystem and the whole BNT.

2. Digital platform for environmental research

To implement digital environmental monitoring of the Baikal natural territory, the ISDCT SB RAS is developing a digital platform (DP) [1-3] (Fig.1), which unites researchers from scientific institutes, employees of departmental organizations and provides them with support for integrated analysis of monitoring data, forecasting and development of recommendations for the conservation of Lake Baikal and adjacent territories. Modern approaches to the creation of digital platforms [4, 5] are usually based on the use of service-oriented architecture (SOA) [6]. This architecture allows to implement a unified information environment supporting algorithmic mutually beneficial relationships of a significant number of independent participants of any process (e.g., environmental monitoring). Within SOA, each participant provides the results of its activity in the form of services, which allows instant access to data and processing methods without the need for coordination. SOA significantly speeds up research. To unify services in the digital platform, the Web Processing Service (WPS) standard being developed by the Open Geospatial Consortium [7] is used. WPS regulates the service invocation protocol and metadata exchange. One of the key entities of SOA is service composition [8], which defines a

computational process in which services of different collectives participate, using control constructs such as sequencing, branching and loops. Accordingly, integrating the work of different collectives to solve complex problems based on a digital platform leads to a variety of service compositions. The digital platform enables the development and application of service compositions, which increases the level of automation of problem solving, from data entry to publication of results.

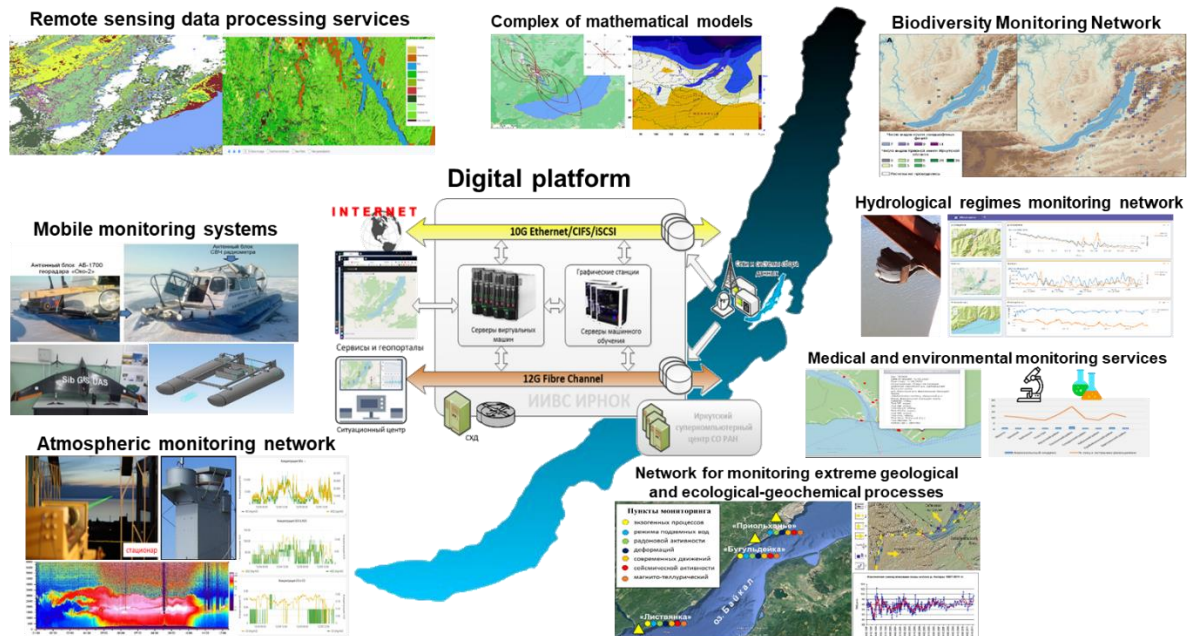


Figure 1: Digital platform for environmental research

The digital platform developed under the project acts as an integrating framework for digital environmental monitoring, is an environment for accumulation, exchange and management of data in a structured form, as well as a system for calling business functions with services of participants of the digital ecosystem (monitoring participants) connected to it through technological interfaces. The digital platform of environmental monitoring consists of three basic components: instrumental, infrastructural and application digital platforms. The Instrumental Digital Platform (IDP) provides access to the participants of the digital ecosystem of monitoring to the development and debugging of applied information and software-hardware, providing them with tool services for processing spatial and temporal data and their interfaces. The IDP includes tools for creating three main types of components embedded in a common environment:

1. development of WPS services and their compositions designed for software data processing, realizing software interoperability standards;
2. creation of data exchange services that provide and make available thematic data collection;
3. deployment of geoportals providing access to thematic data and services.

The IDP contains tools for creating generic geoportals and services in a cloud data center environment. Creation is based on pre-configured virtual machine templates. For WPS services there are templates with Zoo project [9], 52°North Web Processing Service systems that simplify realization of WPS standard.

The Infrastructure Digital Platform (InfDP) provides the creation of applied hardware and software monitoring tools, thematic WPS services of data processing and distributed data storage based on a high-performance information and analytical environment, service-oriented and end-to-end technologies. The InfDP contains means of data delivery, processing, storage, as well as InfDP deployment infrastructure, services (software solutions based on PPI), means of development, debugging and integration of services. In order to solve the project tasks, a computing infrastructure is deployed on the basis of the IICN IREC [10], which provides all project participants with the ability to store large volumes of multiformat spatial and temporal data, high-speed access to them and the ability

to perform high-loaded calculations. The infrastructure consists of a 10G-Ethernet-based high-speed data transmission network connected storage system based on Infortrend EonStor GS 1000 Gen2 (736 TB of capacity), high-performance storage system based on HPE MSA 2060 (27 TB of capacity), high-speed storage area network (SAN) operating on the basis of Fibre Channel 12G protocol, computing cluster of 15 servers with NVidia RTX- 3080, 3090 GPUs (for machine learning). On the basis of the main infrastructure components, using VMWare vSphere 7 virtualization system, a cloud platform was formed, consolidating network and computing resources with resources of data storage systems (network) and flexibly providing them for all systems and services of the project.

The applied digital platform (ADP) operates processed data at the level of a separate group or type of monitoring as a whole, as well as supports the algorithmic exchange of services between independent participants of the digital ecosystem using a unified information and analytical environment and technological infrastructure. The ADP (Fig. 2) consists of a digital monitoring geoportal, data provision services and thematic WPS services.

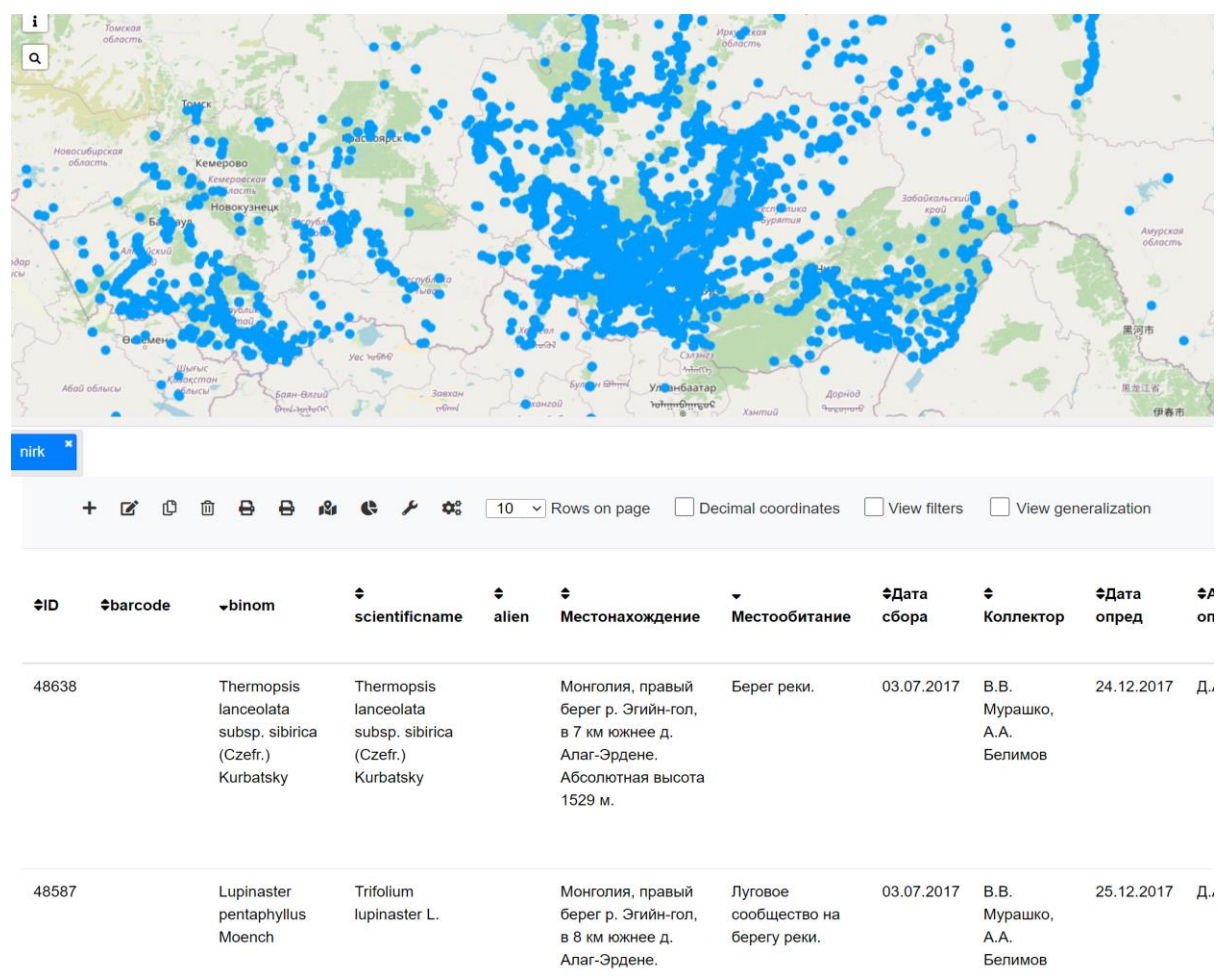


Figure 2: Example of applied digital platform

3. Monitoring technologies based on Earth remote sensing data

The project pays special attention to monitoring methods based on Earth remote sensing (ERS) data processing. A considerable amount of information has been collected: Sentinel-2 satellite images of the Irkutsk Region and the Republic of Buryatia for three years occupy more than 57 terabytes (21 thousand images). To organize a convenient search for space images and quick access to them, a remote sensing data catalog was created, a possibility to display images on the map and support for acquisition of space images of the Earth's surface from various sources was implemented. The images are stored in IICN

IREC storage, their registration is performed by the developed data acquisition components. Image search with regard to sensor, cloud cover, position and date is implemented.

To organize environmental monitoring in large areas, a technology and Web-service (Fig. 3) for classification of multispectral space images Sentinel-2 has been developed [11]. Classification of space images is carried out using the ResNet50 neural network. The technology allows to define a set of classes (surface types), prepare a training sample, verify and classify a large number of images. The training sample is generated using the Web interface based on Sentinel-2 satellite images and field surveys. As input data, the neural network accepts 13 Sentinel-2 channels and a channel built on the basis of local binary patterns to account for textural characteristics. Characteristic feature classes are identified for the BPT: Bogs, Bare Rock, Grassland, Clearings, Transitional Forest/shrubs, Coniferous Forest, Riparian Forest, Deciduous Forest, Mixed Forest, Clouds, Cropland, Residential Area, and Water. 134 space images of BNT were marked, the number of polygonal objects - 2246 pieces, the area of more than 5900 square kilometers. The average classification accuracy amounted to 95%. The developed Web-service allows to increase the efficiency of monitoring of large territories. Classification technology can be used to solve actual problems of large natural areas, in particular, to analyze changes in the forest fund (logging, fires, restoration dynamics), assess the impact of climate change on the landscape, analyze the dynamics of development, inventory of agricultural land, etc.



Figure 3: Example of the result of space images classification

4. Web services for environmental monitoring

The project has created a number of services that solve demanded tasks for the environmental monitoring of BNT. All services are implemented using the components of the instrumental digital platform and support the developed interfaces and data exchange services, which allows simplifying the integration of services and solving complex tasks. The developed services are embedded in the CP and can be applied by users for different tasks. Transaction costs, in this case time costs, are reduced due to the allocation of computing resources that are in constant access, reduction of human participation, standardization of the protocol of service invocation, etc. The created services realize modern methods of data processing and provide unique data of ecological monitoring of BNT. The following thematic services were developed to support digital environmental monitoring of the BNT and Lake Baikal:

- Monitoring service based on remote sensing data processing;
- Biodiversity monitoring service;
- Service for monitoring of hazardous geological processes;
- Atmospheric pollution monitoring service based on data from reference stations;

- Atmospheric pollution monitoring service based on Sentinel-5P satellite data (Fig.4);
- Thunderstorm activity monitoring service;
- Service for monitoring pollution of aquatic ecosystems;
- Service of forest fire risk analysis and forecasting;
- Service for registration of bites by ixodal ticks;
- Service for assessing the impact of forest fires on respiratory organs;
- Service of identification and assessment of changes in the state of vegetation cover;
- Service "Gas and aerosol impurities".
- Service "Generalized data of automatic atmospheric control stations"
- Service "Concentration of substances affecting respiratory organs"; Service "Concentration of substances affecting respiratory organs"
- Service "Atmospheric pollution index";
- Service "Catalog of earthquakes" (Fig.5);
- Service "Fault Field";
- Service "Forest-pathological monitoring";
- Service "The burning of the forest";
- Service "Felling (forest)";
- Service "Phthalates content in water of Selenga River delta channels";
- Service "Level of pH (degree of acidity) in water by seasons";
- Service "Water turbidity by seasons of the year";
- Service "Dissolved oxygen content in water by seasons of the year";
- Service "Content of heavy metals in bottom sediments by seasons of the year";
- Service "Content of heavy metals in water by seasons of the year".

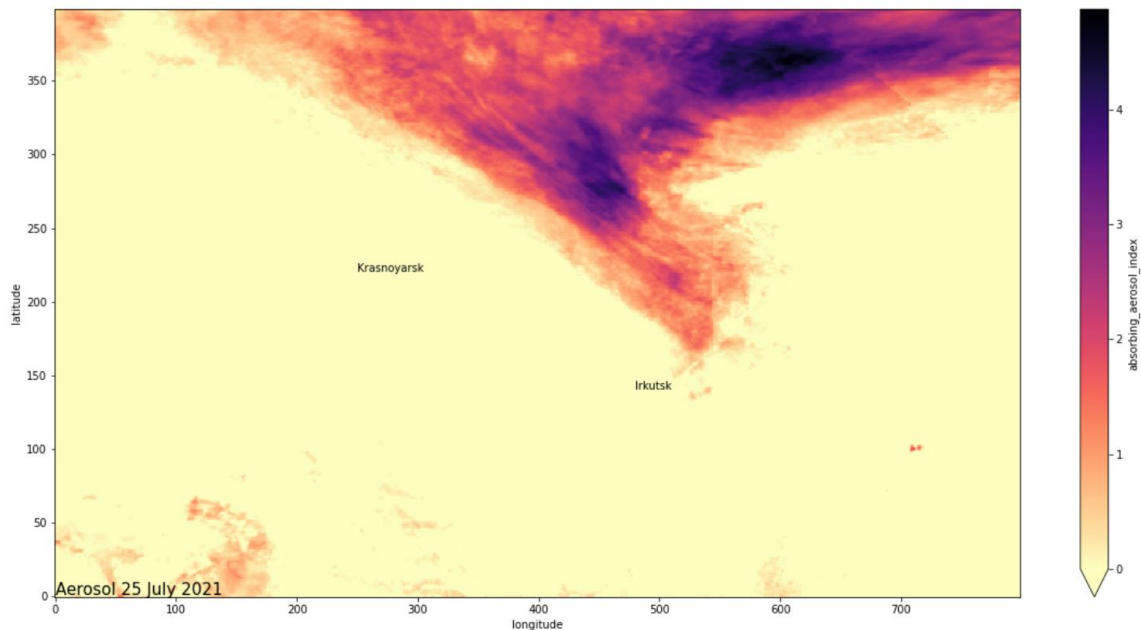


Figure 4: Atmospheric pollution monitoring service

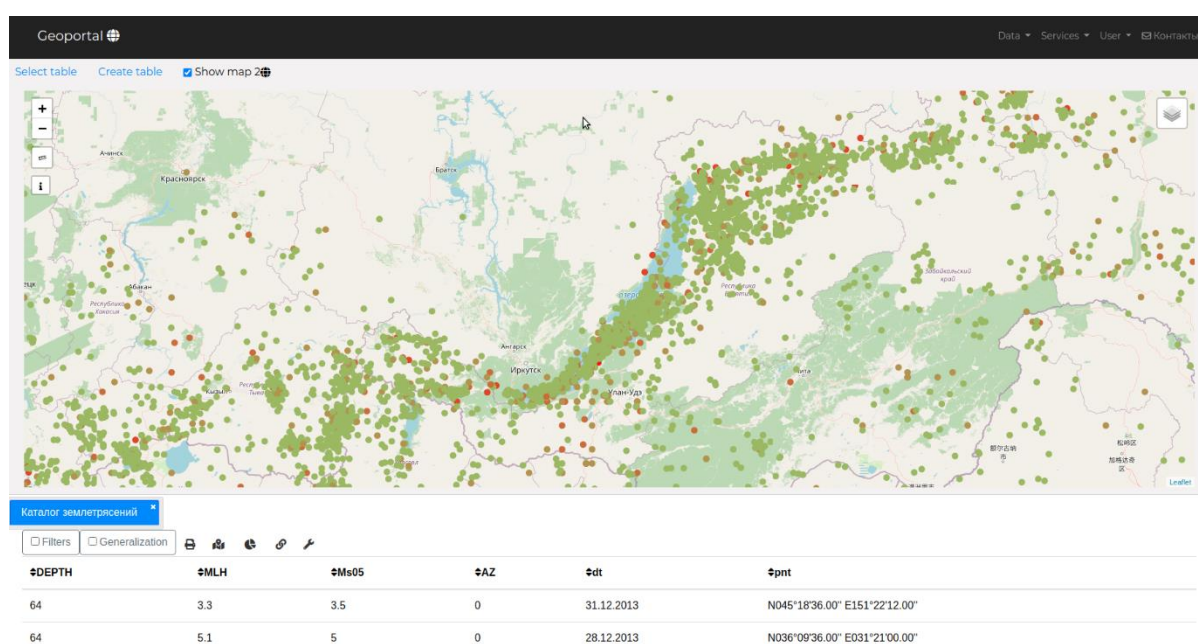


Figure 5: Catalog of earthquakes

5. Conclusion

The results obtained within the framework of the project are primarily of fundamental nature and open new opportunities in the field of development of digital technologies to support large-scale studies in large areas, processing of remote sensing data, mathematical modeling of environmental processes, monitoring and forecasting of natural and anthropogenic emissions in the atmosphere, monitoring and forecasting of hydrological regimes of water bodies, monitoring of dangerous geological processes, monitoring of biodiversity, medical and scientific research, and monitoring of biodiversity. The complex of program components is actively used in practice. More than 200 data services and more than 40 data processing services have been created. 5 different geoportals oriented to different subject areas and teams were deployed. Based on the created services, service compositions were formed, combining data services, processing and publishing services. The created digital platform for environmental monitoring of BNT is an environment for accumulation, exchange and management of data in a structured form, as well as provides the call of analytical functions with the services of participants of the digital ecosystem (monitoring participants) connected to it through technological interfaces.

Acknowledgment:

The work was supported by the Ministry of Science and Higher Education of the Russian Federation, the grant No. 075-15-2024-533 for implementation of Major scientific projects on priority areas of scientific and technological development (the project «Fundamental research of the Baikal natural territory based on a system of interconnected basic methods, models, neural networks and a digital platform for environmental monitoring of the environment»).

References:

- [1] Bychkov I.V., Fedorov R.K., Fereferov E.S. Instrumental components of the digital platform for environmental monitoring of the Baikal natural territory // Computational Technologies. 2023. Vol. 28, №6. P. 95-107. doi: 10.25743/ICT.2023.28.6.009.
- [2] Bychkov I.V., Fereferov E.S. Digital Technologies for Monitoring and Forecasting the Environmental Situation in Siberia //Herald of the Russian Academy of Sciences, 2022, Vol. 92, No. 2, pp. 133–140. doi: 10.1134/S101933162202006X.

- [3] Бычков И.В., Гладкочуб Д.П., Ружников Г.М. и др. Фундаментальные основы, методы и технологии цифрового мониторинга и прогнозирования экологической обстановки Байкальской природной территории. 2022. Новосибирск: СО РАН, 2022. 345 с.
- [4] De Reuver M, Sørensen C and Basole R C 2018 The digital platform: A research agenda J. Inf. Technol. 33 124-35. doi: 10.1057/s41265-016-0033-3
- [5] Osipov Yu M, Yudina T N and Geliskhanov I Z Digital platform as an institution of the era of technological breakthrough // Economic Strategies. 2018. 5(155) P. 22-29
- [6] Bih, J. Service oriented architecture (SOA) a new paradigm to implement dynamic e-business solutions // Ubiquity. 2006. Vol. 4. P. 1–17.
- [7] WPS - Standardized web-based geo-processing. URL: <https://52north.org/software/software-projects/wps/>
- [8] Edmond, D. & ter Hofstede, A. (2000), Service composition for electronic commerce, in 'Proceedings of the Pacific Asia Conference on Information Systems(PACIS-2000)', Hong Kong.
- [9] ZOO-Project – Open WPS platform. URL: <http://www.zoo-project.org/>
- [10] The Integrated information and computing network of the Irkutsk scientific and educational complex. URL: <http://net.icc.ru/>
- [11] Bychkov I.V., Ruzhnikov G.M., Fedorov R.K., Popova A.K., Avramenko Y.V. Classification of Sentinel-2 satellite images of the Baikal Natural Territory // Computer Optics. 2022. Vol. 46, №1. pp. 90-96. doi: 10.18287/2412-6179-CO-1216.

Reducing Manual Labeling Effort by Identifying the Most Informative Unlabeled Data via Active Learning

Velibor Ilić¹

¹ The Institute for Artificial Intelligence Research and Development of Serbia - IVI, Fruškogorska 1, 21000 Novi Sad, Serbia

velibor.ilic@ivi.ac.rs

Abstract:

This paper presents a novel approach to reducing manual labeling effort by identifying and prioritizing the most informative samples from large unlabeled datasets through active learning frameworks. We introduce two specialized frameworks, ALSCN (Active Learning using a Self-Correcting Neural Network) and ALWOD (Active Learning for Weakly-Supervised Object Detection), designed for classification and object detection tasks, respectively. ALSCN leverages selective sampling to achieve high classification accuracy with minimal labeled data, while ALWOD combines active learning with semi- and weakly supervised paradigms, using a student-teacher model setup to enhance detection performance. Evaluated on benchmark datasets, with ALSCN on MNIST, Fashion MNIST, and CIFAR-10, and ALWOD on COCO2014, VOC2007, and RealPizza10, both frameworks outperform traditional methods by significantly reducing annotation needs while maintaining competitive accuracy. These two active learning frameworks demonstrate a substantial reduction in manual labeling effort, achieving high accuracy with minimal annotated data.

Keywords:

Active learning, machine learning, data labeling, classification, Object Detection

1. Introduction

In the field of machine learning, labeled data serves as the foundational backbone for training supervised models. High-quality labeled datasets enable algorithms to learn input-output mappings, recognize patterns, and make accurate predictions on new, unseen data. However, the process of obtaining labeled data is often fraught with challenges, particularly when scaling to large datasets [1]. Manual labeling is a labor-intensive and time-consuming endeavor that requires significant human resources and expertise. As data sets grow in size, the effort and cost associated with labeling each data point becomes greater. This scalability issue is further complicated by the need for precise and consistent labeling to ensure model accuracy, which can be difficult to maintain across large teams of annotators. Moreover, manual labeling is susceptible to human error and subjectivity, potentially introducing inconsistencies that degrade model performance. Consequently, the reliance on extensive manual labeling poses a substantial bottleneck in the development and deployment of effective machine learning models, especially in domains where data is abundant, but labels are scarce or expensive to obtain [2].

Active learning offers an effective way to reduce the challenges of manual labeling and make data annotation faster and easier [3]. It is a machine learning paradigm where the model proactively selects the most informative unlabeled instances for labeling, rather than relying on random sampling. By identifying and prioritizing data points that are expected to contribute the most to the learning process, active learning aims to achieve high model performance with significantly fewer labeled examples. This targeted approach reduces the overall labeling effort by focusing human annotation on data that resolves the model's uncertainties or addresses its weaknesses. As a result, active learning not only accelerates the training process but also optimizes resource allocation by minimizing redundant or uninformative labeling tasks. Implementing active learning techniques can therefore substantially decrease the cost and time associated with manual labeling, making it a valuable tool for handling large datasets and improving the scalability of machine learning applications [4, 5].

Current labeling processes are inefficient because they depend heavily on manual annotation, which is both time-consuming and resource intensive. As datasets become larger and more complex, these

inefficiencies scale up, leading to increased costs and delays in model development. Therefore, there is a need for methods that can reduce manual effort while maintaining, or even improving model performance to make machine learning more scalable and practical [6,7].

The primary objective of this paper is to develop efficient frameworks [8] for reducing the manual labeling effort required in training machine learning models on large datasets. This is achieved by identifying the most informative unlabeled data that, when labeled, significantly enhance model performance. The goal is to minimize redundant labeling without compromising, and potentially even improving, the accuracy of the resulting models compared to those trained on fully labeled datasets.

This research introduces two active learning frameworks, Active Learning using a Self-Correcting Neural Network (ALSCN) [9] and Active Learning for Weakly-Supervised Object Detection (ALWOD) [10], each designed to improve data labeling efficiency and model performance in distinct settings. ALSCN optimizes label selection from large unlabeled datasets to enhance classification accuracy, while ALWOD combines active learning with weak and semi-supervised methods [11, 12, 13, 14, 15] to streamline annotation in object detection [16, 17], effectively narrowing the gap between partially and fully labeled data models.

2. Proposed Methodology

Active learning is a machine learning paradigm aimed at reducing manual labeling efforts. The active learning strategy involves selecting the most uncertain or informative data points from a large pool of unlabeled examples to maximize model improvement with minimal labeling effort. By prioritizing these high-value samples, the model learns more efficiently, reducing the need for extensive manual annotations while achieving robust performance. This iterative process of selecting, annotating, and retraining continues until the model reaches a desired level of accuracy with fewer labeled data points. Common techniques include **uncertainty sampling** [18], where the model selects data points it is least confident about, **query by committee** [19], which involves multiple models voting on data points with the most disagreement, and **diversity sampling** [20], ensuring a wide representation of the dataset is included. These methods help improve model performance while requiring fewer labeled instances.

The annotation workflow starts by identifying the most informative data samples through criteria such as uncertainty or model disagreement, ensuring that only high-priority items are selected for review. Human annotators then concentrate on these chosen samples, either providing detailed labels or validating model-generated predictions, significantly reducing the time and effort required for annotation. This labeled data is subsequently integrated back into the model, enhancing its accuracy and effectiveness as it learns from each new set of valuable examples.

The label correction mechanism identifies potential labeling errors in the dataset by comparing model predictions with existing labels, flagging discrepancies that may indicate inaccurate annotations. Human annotators review these flagged instances, correcting any detected errors to improve the quality and reliability of the labeled data. This corrected data is then reintegrated into the training process, leading to a more accurate model and reducing the impact of incorrect labels over time.

Self-correcting neural network (SCN) is neural network architecture or learning algorithm designed to identify and correct their own errors during training or inference. The core concept involves enhancing the model's ability to recognize instances where its predictions might be incorrect or uncertain and then adjust its parameters or outputs accordingly. This self-correction mechanism aims to improve the overall accuracy and robustness of the model, particularly when dealing with noisy data or limited labeled datasets [21].

Concept of Self-Correcting Neural Networks typically incorporate mechanisms such as:

- **Error Detection:** The network monitors its predictions to detect potential errors, often by assessing confidence levels or prediction uncertainties. Low-confidence predictions may indicate areas where the model lacks sufficient knowledge [22, 23].
- **Feedback Loops:** Upon detecting potential errors, SCNs utilize feedback mechanisms to adjust internal parameters. This can involve reweighting certain data points, modifying learning rates, or updating weights and biases to correct mispredictions [12].

- **Iterative Refinement:** The model undergoes multiple iterations of prediction and adjustment, refining its outputs each time based on the errors identified in previous iterations.

In prior research, SCNs have been employed across various domains to enhance model performance and data efficiency:

- **Noise Reduction and Label Correction:** SCNs have been used to detect and correct mislabeled or noisy data within training datasets. By identifying inconsistencies between predicted and actual labels, the network can flag potentially incorrect annotations for review or automatically adjust them, leading to cleaner training data and improved learning outcomes [24].
- **Active Learning Integration:** SCNs have been integrated into active learning frameworks to select the most informative unlabeled instances for annotation. By self-assessing prediction uncertainties, the network can determine which data points would most benefit from manual labeling, thereby reducing the overall labeling effort while maximizing model improvement.
- **Anomaly Detection:** In fields like anomaly detection, SCNs can identify deviations from normal patterns and adjust their parameters to better capture rare or unexpected events, enhancing the model's sensitivity and specificity.

Weakly supervised object detection aims to train models using limited or imprecise annotations, such as image-level labels instead of detailed bounding boxes [25]. Common approaches in this domain include multiple instance learning, where the model learns to associate image-level labels with specific regions within an image to localize objects. Techniques like class activation mapping are used to highlight regions of interest that correspond to certain classes. Semi-supervised learning for object detection leverages a small set of labeled data alongside a larger pool of unlabeled data. Methods such as pseudo-labeling involve the model generating labels for unlabeled images based on its current predictions, which are then used to retrain the model. Consistency regularization techniques encourage the model to produce stable predictions under various perturbations or augmentations of the input data, thereby improving its generalization capabilities.

3. ALSCN and ALWOD frameworks

This research introduces two complementary active learning frameworks, ALSCN and ALWOD, each crafted to enhance data labeling efficiency and boost model performance across distinct tasks. ALSCN is tailored for image classification tasks with large pools of unlabeled data, where it iteratively selects the most informative examples to be labeled by class, reducing the amount of data requiring manual annotation while consistently improving model accuracy. This selective approach not only conserves labeling resources but also enables models to generalize more effectively with fewer labeled samples.

ALWOD framework tackles the challenges in object detection, where limited training data with precise labels is often a constraint. By combining active learning with weak and semi-supervised methods, ALWOD uses a small set of fully labeled images and a larger set of weakly tagged images to “warm-start” training. Its unique acquisition function leverages student-teacher model [26] disagreement and uncertainty to select the most informative images for annotation, making the labeling process faster and more accurate.

3.1. ALSCN Algorithm Details

Figure 1 illustrates the ALSCN algorithm, showcasing its process of iteratively selecting data samples for labeling to enhance model performance efficiently.

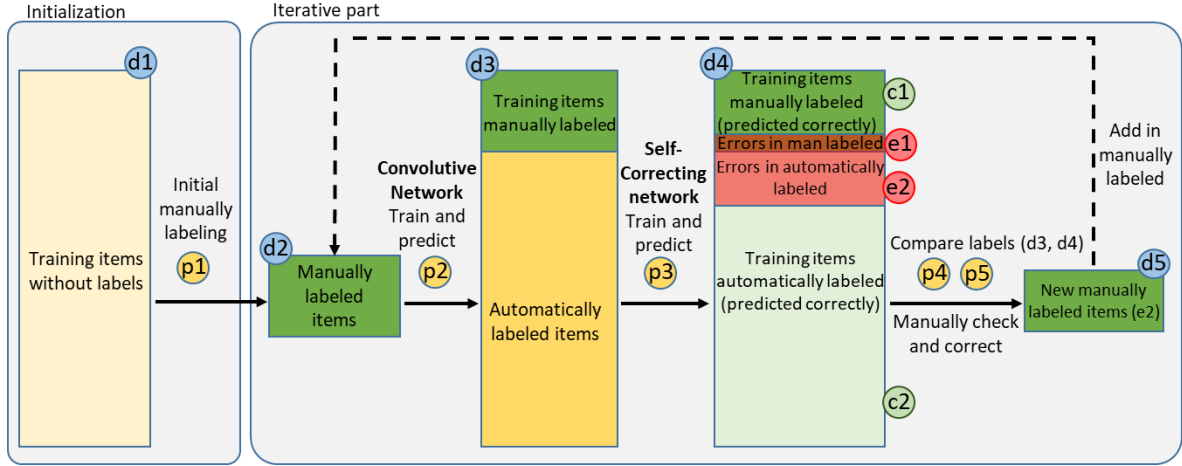


Figure 1: Active Learning using a Self-Correcting Neural Network (ALSCN)

In the initialization phase, we start with an unlabeled dataset (d_1) and randomly select a small subset for manual labeling (p_1), creating the initial labeled dataset (d_2). In the iterative phase, a network (p_2) is trained on (d_2) and used to predict labels for the remaining unlabeled items, generating a fully labeled dataset (d_3). A self-correcting network (p_3) is then trained on (d_3), and its predictions on the entire dataset create a refined dataset (d_4). A simple comparison between (d_3) and (d_4) allows us to categorize the items into four groups. If errors appear in the group of previously manually labeled items (e_1), it may indicate that the initial items were incorrectly labeled. Such items should be selected for double-checking by a human expert to ensure data quality.

Errors in automatically labeled items (e_2) are even more interesting because these items have not been reviewed by humans before. Therefore, all items in this group should be manually checked and corrected if necessary. After this manual verification, these items should be added to the dataset containing manually labeled items (d_2).

This process represents the final step in the iterative part of our algorithm, and the previously described steps are repeated in a loop. We can monitor the number of errors in the automatically labeled items (e_2) that need correction and use this count as a condition for terminating the loop. If the number of such errors falls below a certain threshold or if satisfactory results have been achieved, we can consider that sufficient items from the training set have been selected for manual labeling.

Furthermore, the networks (p_2) and (p_3) can load previously saved network parameters, which reduces training time and increases accuracy. Since the networks trained in the previous iteration can already accurately predict the labels for most items in the new dataset of the current iteration, we only need to fine-tune them to learn the differences from the previous iteration rather than training them from scratch. This is especially evident in later iterations when only a few items are detected per iteration.

In relation to the previous iteration, the dataset (d_2) is extended with new manually labeled items, and in dataset (d_3), some automatically labeled items are replaced with these new manually labeled items. This iterative refinement continues until the model reaches the desired level of accuracy with minimal manual labeling effort.

3.2. ALWOD Algorithm Details

Figure 2 illustrates the ALWOD algorithm, highlighting its fusion of active learning with weakly and semi-supervised methods to effectively select and annotate key images for object detection tasks.

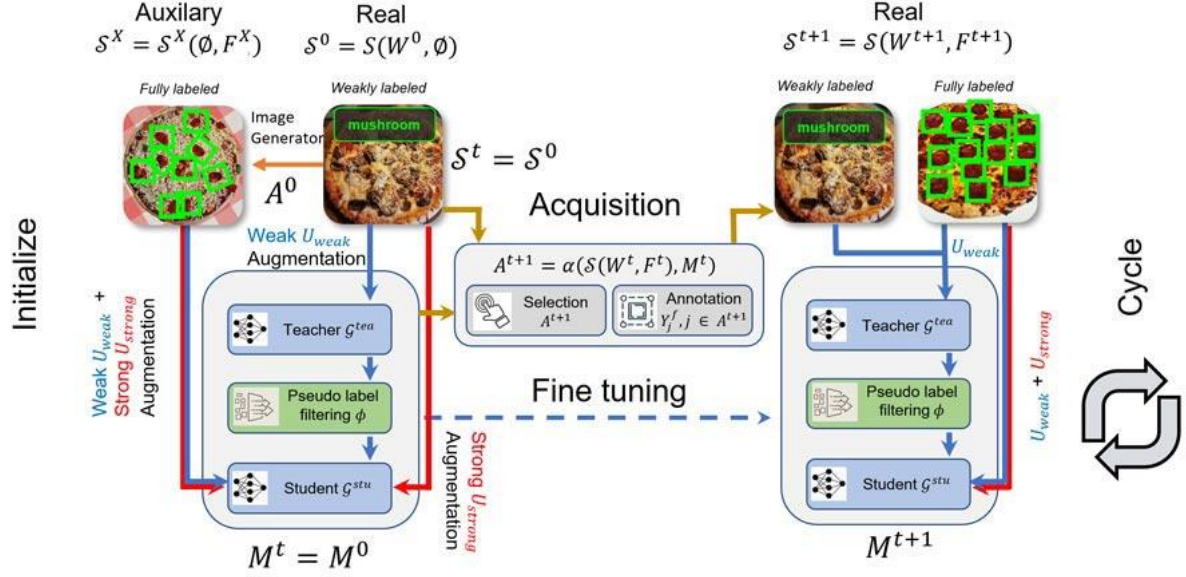


Figure 2: Active Learning for Weakly-Supervised Object Detection (ALWOD)

ALWOD is an active learning framework designed for object detection, combining the strengths of active learning with semi-supervised and weakly supervised paradigms. It uses a student-teacher network setup, where both networks are initialized on a minimal labeled dataset and a large weakly labeled set. Through iterative learning cycles, ALWOD strategically selects the images for annotation, enhancing detection accuracy with reduced labeling effort.

In the initial learning phase, ALWOD starts with a weakly labeled set of real-world images, S^x , without fully annotated bounding boxes, which poses a challenge for accurate object detection. To overcome this, the framework uses an auxiliary fully-labeled synthetic dataset, S^x , composed of real-world backgrounds with realistic foreground objects, as a warm-start for training. This dataset provides a solid initial training base, allowing the student-teacher network to develop an effective detection model.

The auxiliary set, S^x , is created to enhance the model's initial performance without high annotation costs, using synthetically generated images that combine real-world backgrounds with cropped, realistic objects. This fully labeled set enables the model to learn robust object detection features before applying active learning to real-world data. The use of S^x , alongside weakly labeled data, is crucial for reducing annotation effort while achieving strong baseline performance.

Active Learning Cycle ($t > 0$), during each active learning cycle, S^{t+1} , ALWOD selects a batch of informative images from the weakly annotated data using an acquisition function based on model uncertainty and student-teacher disagreement. These images are then manually annotated, providing full labels that replace the existing weak labels and expand the fully labeled set. This newly labeled data is used to fine-tune the student-teacher model, and the cycle continues until the annotation budget is exhausted or the model's performance converges.

The active learning strategy in ALWOD aims to select the images from dataset within a limited annotation budget. By choosing images that will most reduce model loss and improve detection performance, ALWOD prioritizes samples based on signals from a student-teacher network setup. The two main signals used for selection are model disagreement and prediction uncertainty, both of which help pinpoint images that are most likely to benefit from manual annotation.

Model disagreement measures the variance in predictions between the student and teacher networks, with the teacher network acting as a moving average of the student models. Ideally, the predictions of the student network would align with the teacher's; however, large discrepancies signal that an image may be challenging and worthy of manual review. Disagreement is quantified by calculating per-class average precision differences between the student's predictions and those of the teacher, identifying images where the models show significant predictive differences.

Image uncertainty is determined by the entropy of the teacher network's predictions for each detected object in an image, with higher entropy indicating more uncertainty in the model's

classifications. This signal aggregates uncertainties across all predicted objects in an image, flagging images with high uncertainty scores as candidates for manual labeling. Images with the highest entropy scores are selected because the model's predictions are less confident, suggesting that manual annotation could improve training data quality.

The acquisition function combines the model disagreement and image uncertainty scores to rank images for selection, prioritizing those with the highest product of these scores. By fusing these two signals, the function identifies images that are both challenging and uncertain, likely requiring human annotation to improve model performance. Each cycle of active learning selects a new batch of these high-score images, refining the training data while staying within the annotation budget.

The annotation tool developed for ALWOD simplifies object detection labeling by presenting images in an ordered list based on their acquisition scores. Each image comes with pre-generated bounding boxes and class labels from both the student and teacher networks, which undergo non-maximum suppression and confidence thresholding to reduce redundancy. Annotators are tasked with reviewing the suggested bounding boxes by selecting those that overlap with true objects, adjusting bounding box categories, and verifying bounding box quality. For any object missed by the model, annotators can draw new bounding boxes and assign labels. This workflow reduces the time and effort required for accurate annotations, balancing efficiency and precision in the labeling process.

3.3. Datasets for Evaluating the Performance of ALSCN and ALWOD Frameworks

The ALSCN framework is evaluated on three widely recognized datasets - MNIST, Fashion MNIST, and CIFAR-10 (Figure 3) - which provide a variety of visual classification challenges ranging from handwritten digits to complex object classes. These datasets enable the assessment of ALSCN's ability to enhance model accuracy with minimal labeled data.



Figure 3: Data sets used for testing performance of ALSCN framework

In contrast, ALWOD is tested on COCO2014, VOC2007, and RealPizza10 (Figure 4), which are benchmark datasets for object detection, offering complex scenes with multiple objects and variable annotations. These diverse datasets demonstrate ALWOD's effectiveness in reducing annotation effort while achieving robust detection performance in real-world applications.



Figure 4: Data sets used for testing performance of ALWOD framework

3.4. Results of ALSCN framework

To assess the method's effectiveness, we trained networks with identical architectures on three datasets—MNIST, Fashion MNIST, and CIFAR-10—using both manually labeled items and randomly selected items for comparison. Figure 5 shows the prediction accuracy across different training sample sizes: the red line represents accuracy of a CNN trained only on manually labeled items, the green line

shows a self-correcting network trained on both manually and automatically labeled items, and the blue line represents the average accuracy of five networks trained on random selections.

The results highlight that, with our algorithm, only a fraction of each dataset was needed to achieve comparable or superior accuracy: 6.11% for MNIST, 23.92% for Fashion MNIST, and 59.4% for CIFAR-10, showing substantial reduction in manual labeling effort while maintaining high performance.

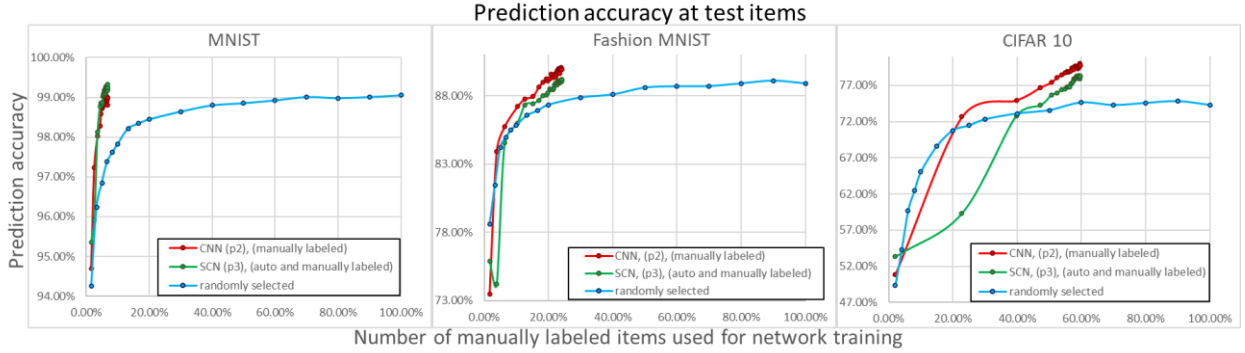


Figure 5: ALSCN results show the accuracy of predictions based on the number of items used to train the neural network

3.5. Results of ALWOD framework

ALWOD outperforms a range of acquisition strategies, including image uncertainty, model disagreement, loss, entropy-sum, random sampling, core-set, and BiB [27], across VOC2007 [28], RealPizza10 [29], and COCO2014 [30]. With five annotation cycles and a fixed image budget per cycle, ALWOD's product-based acquisition function demonstrates clear advantages, particularly noticeable from cycle 3 onwards, achieving higher accuracy than competing methods. This improvement leverages the unique student-teacher network disagreement approach in ALWOD, enabling it to consistently select the most informative samples for annotation.

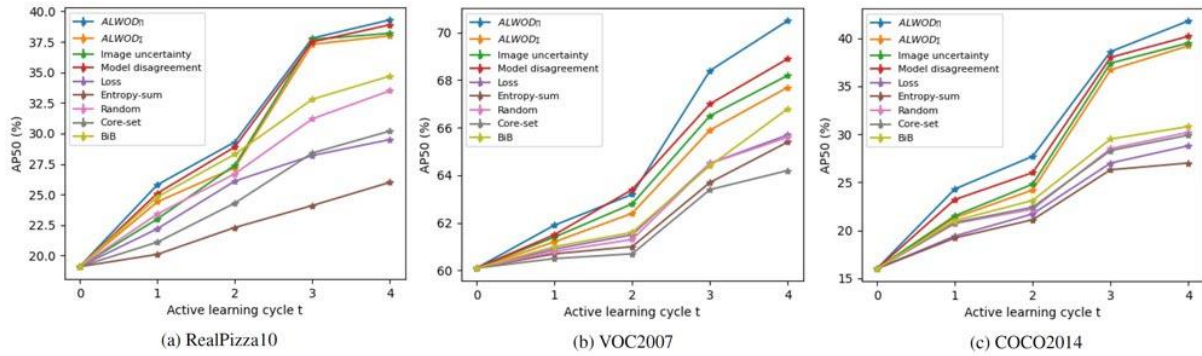


Figure 6: ALWOD results show the accuracy of predictions per sessions

4. Conclusions

This research presents two active learning frameworks, ALSCN and ALWOD, each designed to optimize data labeling and model performance in different environments. ALSCN focuses on iteratively selecting the most informative items from a large pool of unlabeled data, reducing labeling demands while improving accuracy. ALWOD, on the other hand, addresses object detection challenges by combining active learning with weakly and semi-supervised paradigms. By utilizing a minimal labeled set alongside a larger weakly tagged set, ALWOD's unique student-teacher model disagreement and uncertainty metrics guide the selection of key images for annotation, improving the efficiency of the object detection process.

Both ALSCN and ALWOD frameworks demonstrate a clear reduction in manual labeling needs. ALSCN identifies high-value samples, allowing datasets to grow with minimal additional manual effort. ALWOD, through a selective labeling strategy, enables human annotators to validate only specific model-proposed detections, providing an efficient approach to object detection with limited labeled data.

These frameworks advance the fields of machine learning and data annotation by introducing novel approaches to label-efficient training. ALSCN optimizes supervised learning through selective sampling, while ALWOD's innovative active learning loop helps object detectors trained on partially labeled datasets perform comparably to those trained on fully labeled ones. Together, these frameworks promote more sustainable, scalable data annotation practices in complex machine learning tasks.

References:

- [1] Wang, K., Zhang, D., Li, Y., Zhang, R. and Lin, L., (2016) "**Cost-effective active learning for deep image classification**" IEEE Transactions on Circuits and Systems for Video Technology, 27(12), pp.2591-2600.
- [2] Wei, K., Iyer, R. and Bilmes, J., (2015) "**Submodularity in data subset selection and active learning**", In International Conference on Machine Learning (pp. 1954-1963).
- [3] Settles, B., (2009). "**Active learning literature survey**", University of Wisconsin-Madison Department of Computer Sciences.
- [4] Sener, O. and Savarese, S., (2017), "**Active learning for convolutional neural networks: A core-set approach**", International Conference on Learning Representations (ICLR) 2018
- [5] Kim, T., Lee, K., Ham, S., Park, B., Lee, S., Hong, D., Kim, G.B., Kyung, Y.S., Kim, C.S. and Kim, N., (2020), "**Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: Evaluation on kidney segmentation in abdominal CT**", Scientific reports, (Nature Publisher Group), 10(1), pp.1-7.
- [6] Seeger, M., (2001), "**Learning with labeled and unlabeled data**", Institute for Adaptive and Neural Computation, University of Edinburgh
- [7] Goudjil, M., Koudil, M., Bedda, M. and Ghoggali, N., (2018), "**A novel active learning method using SVM for text classification**", International Journal of Automation and Computing, 15(3)
- [8] Ilić V., Prodanović N., Kalušev V. (2023), "**Applying Active Learning in NLP Domain**", AAI2023, The Second Serbian International Conference on Applied Artificial Intelligence (SICAAI)
- [9] Ilić V., Tadić J. M. (2021) "**Active Learning using a Self-Correcting Neural Network (ALSCN)**", Applied Intelligence, The International Journal of Research on Intelligent Systems for Real Life Complex Problems, ISSN: 0924-669X, Springer
- [10] Wang Y., Ilić V., Li J., Kisačanin B., Pavlović V. (2023) "**ALWOD: Active Learning for Weakly-Supervised Object Detection**", The IEEE/CVF Conference on Computer Vision and Pattern Recognition ICCV 2023
- [11] Gao, M., Zhang, Z., Yu, G., Arık, S.Ö., Davis, L.S., and Pfister, T., (2020), "**Consistency-based semi-supervised active learning: Towards minimizing labeling cost**", In European Conference on Computer Vision (pp. 510-526). Springer, Cham.
- [12] Fazakis, N., Vasileios G. K., Christos K. A., Stamatis K., Sotiris K., (2019), "**Combination of Active Learning and Semi-Supervised Learning under a Self-Training Scheme**", Entropy 21, no. 10, 988.
- [13] Chen, T., Kornblith, S., Swersky, K., Norouzi, M. and Hinton, G., (2020), "**Big self-supervised models are strong semi-supervised learners**", Advances in Neural Information Processing Systems, (NeurIPS 2020)
- [14] S. Huang, T. Wang, H. Xiong, J. Huan, and D. Dou, "**Semi-supervised active learning with temporal output discrepancy**", in ICCV, 2021.
- [15] J. Guo, H. Shi, Y. Kang, K. Kuang, S. Tang, Z. Jiang, C. Sun, F. Wu, and Y. Zhuang, "**Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels**", in ICCV, 2021.
- [16] C.-C. Kao, T.-Y. Lee, P. Sen, and M.-Y. Liu, "**Localization-aware active learning for object detection**", in ACCV, 2018.
- [17] J. Wu, J. Chen, and D. Huang, "**Entropy-based active learning for object detection with progressive diversity constraint**" in CVPR, 2022.

- [18] Sharma, M. and Bilgic, M., (2017), “***Evidence-based uncertainty sampling for active learning***”, Data Mining and Knowledge Discovery, 31(1), pp.164-202.
- [19] Grimova, N. and Macas, M., (2019) “***Query-By-Committee Framework Used for Semi-Automatic Sleep Stages Classification***”, In Multidisciplinary Digital Publishing Institute Proceedings (Vol. 31, No. 1, p. 80).
- [20] Yang, Y., Ma, Z., Nie, F., Chang, X. and Hauptmann, A.G., (2015). “***Multi-class active learning by uncertainty sampling with diversity maximization***”, International Journal of Computer Vision, 113(2), pp.113-127.
- [21] Liu, X., Li, S., Kan, M., Shan, S. and Chen, X., (2017), “***Self-error-correcting convolutional neural network for learning with noisy labels***”, In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 111-117). IEEE.
- [22] Neutatz, F., Mahdavi, M. and Abedjan, Z., (2019), “***Ed2: A case for active learning in error detection***”, In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 2249-2252).
- [23] Bekker, A.J. and Goldberger, J., (2016), “***Training deep neural-networks based on unreliable labels***”, In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2682-2686). IEEE.
- [24] Jakramate B., Kaban A., (2012), “***Label-noise robust logistic regression and its applications***”, in Machine Learning and Knowledge Discovery in Databases, pp. 143–158.
- [25] H. V. Vo, O. Simeoni, S. Gidaris, A. Bursuc, P. Perez, and J. Ponce, “***Active learning strategies for weakly-supervised object detection***”, in ECCV, 2022.
- [26] P. Mi, J. Lin, Y. Zhou, Y. Shen, G. Luo, X. Sun, L. Cao, R. Fu, Q. Xu, and R. Ji, “***Active teacher for semi-supervised object detection***”, in CVPR, 2022.
- [27] Huy V Vo, Oriane Simeoni, Spyros Gidaris, Andrei Bursuc, Patrick Perez, and Jean Ponce. ***Active learning strategies for weakly-supervised object detection***. In ECCV, 2022.
- [28] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. ***The pascal visual object classes (VOC) challenge***. IJCV, 2010.
- [29] Yuting Wang, Ricardo Guerrero, and Vladimir Pavlovic. ***D2DF2WOD: Learning object proposals for weakly supervised object detection via progressive domain adaptation***. In WACV, 2023.
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. ***Microsoft COCO: Common objects in context***. In ECCV, 2014.

AIIT 2024

REGULAR PAPERS



Malware and Social Engineering Threats

Aleksandar Bulajic¹

¹ PORT85, Buddingevej 48, 2800 Kongens Lyngby, Denmark

mail@port85.com

Abstract:

Cyber threats, including fraud, scams, ransomware, and espionage, have become widespread, endangering both individuals and organizations. The widespread availability of digital technologies and networks provides an environment conducive to illegal activities, leaving no one immune to potential harm. Cyber attackers primarily use Social Engineering techniques to breach systems, often deploying malicious code. The situation is worsened by the availability of Do-It-Yourself (DIY) kits, which enable the easy configuration and dissemination of malicious software at minimal or no cost. The emergence of Artificial Intelligence (AI) and Generative AI (GenAI) further complicates cybersecurity, with the potential to intensify these threats. Although there is hope that AI may bolster defenses, the history of cyber breaches suggests that malicious actors often remain one step ahead. Complex international dynamics and hybrid warfare in cyberspace present additional dangers. This paper explores recent statistics and trends in malicious software and cybercrime techniques, with a particular focus on Social Engineering-based attacks.

Keywords:

Cybercrime, fraud, identity theft, ransomware, Social Engineering, phishing, smishing, vishing, pharming.

1. Introduction

Major entities such as the NSA, Microsoft, Facebook, LinkedIn, and First American Financial Corporation have suffered significant losses due to cybercrime, emphasizing the widespread nature of the threat. The history of successful cyberattacks, including those targeting Microsoft and Mærsk A/S, illustrates the enormous financial and operational damage caused by such incidents.

Data breaches have impacted numerous companies, such as Facebook, LinkedIn, and JPMorgan Chase, resulting in the exposure of sensitive information. Meanwhile, individual users often fall victim to mass attacks, disseminated through email, messaging platforms, and social networks, aiming to compromise systems or steal personal data.

The first part of this paper highlights a small number of the most notable cybercrime attacks, provides the latest statistics on the financial toll of cyberattacks, and examines which countries and industries are most affected.

The second part covers the common types of malwares and provides statistics on the most prevalent malware trends in cyberattacks.

The third part focuses on Social Engineering, a widely used technique that manipulates individuals into actions that compromise cybersecurity.

The fourth part briefly discusses Artificial Intelligence (AI) and the challenges posed to cybersecurity by AI and Generative AI (GenAI).

1.1. Short history of successful Cyber Crime Attacks

In 2021, Microsoft became a victim of one of the largest cyberattacks in U.S. history, during which hackers took control of over 30,000 business servers for three months and gained access to various business and government emails [1].

In 2017, Maersk A/S, one of the world's largest shipping companies, suffered a cyberattack involving Petya malware, which infected 49,000 servers, 2,500 applications, and impacted 80% of A.P.

Møller-Mærsk. This incident cost Maersk A/S nearly two billion Danish crowns, approximately \$290 million USD, as of May 2024 [2]. Petya malware encrypts the Windows file system and master boot record, preventing Windows servers from booting. It can also demand a Bitcoin payment for system access. Below is an image that Petya malware can display as part of its payload [3]:



Figure 1: Petya-malware payload picture

In 2021, a Facebook data breach exposed the personal information of 530 million users, including names, phone numbers, passwords, and account names. This breach was one of several that Facebook experienced since 2013. In 2019, an investigation revealed that a Vietnamese hacker group had compromised 300 million Facebook accounts. Among the most notorious incidents was the Cambridge Analytica scandal, where data from 50 to 90 million Facebook users was sold [1].

The list of data breaches is extensive and includes LinkedIn, Instagram, JPMorgan Chase, eBay, Amazon, Tesla, and First American Financial Corporation, which suffered a leak of 885 million file records containing sensitive data such as bank account numbers, bank statements, Social Security numbers, and driver's licenses [1].

A simple Google search regarding data breaches at the U.S. National Security Agency (NSA) reveals numerous cases, with the Edward Snowden incident being the most notable. Below is a snapshot from a Google search:

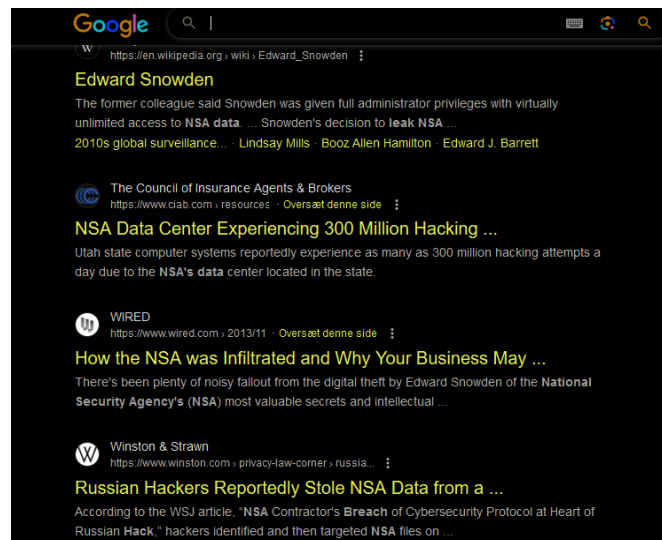


Figure 2: Google search snapshot

VPNCentral conducted a study on the most affected data breach accounts on popular internet sites, finding that Facebook is the most impacted, followed by Instagram and Spotify [4].

Ordinary individuals are rarely direct targets; instead, cybercriminals often employ bulk attacks. These attacks aim to reach a large number of people, banking on the chance that someone will click on an infected link and install malware on their computer. Such attacks can be disseminated globally via email, messaging systems, social networks, and file-sharing networks. When a user clicks on a link in an infected message, it may install malware, steal usernames and passwords, or redirect the user to compromised sites to collect sensitive information such as credit card numbers or Social Security numbers. This information can be used for illegal activities or traded on dark web platforms like Tor and IP4.

1.2. Cost of Cyber Attack

The costs associated with cyberattacks can be staggering, encompassing data loss, downtime, reputational damage, and legal repercussions. Organizations that fail to report breaches promptly often incur additional expenses and experience prolonged breach lifecycles. The healthcare industry faces the highest costs related to data breaches, followed by the financial, pharmaceutical, and energy sectors. The following image illustrates the average costs of a successful cyberattack, based on IBM research conducted in 2023 [5]:



Figure 3: The global average cost of data breach in 2023 [4].

The use of security AI and automation can help reduce the costs associated with data breaches [5]. However, these costs can still be substantial, and recovering fully to resume normal operations can take time. Reputational damage and downtime can lead to customer loss and negatively impact future revenue.

According to IBM research, 37% of organizations fail to report ransomware attacks to the appropriate legal authorities, which can result in an additional cost of \$470,000. These organizations pay 9.6% more and experience a breach lifecycle that is 33 days longer than those that report breaches properly [5].

Only one-third of all data breaches are identified by organizations; the remainder are reported by attackers or third-party sources [5]. If an attacker discloses a breach, it can cost about \$1 million more than if the breach is discovered internally [5].

The healthcare industry reports the highest costs associated with data breaches, followed by the financial, pharmaceutical, energy, and technology sectors [5].

1.3. Top affected countries and industries

According to IBM [5], the top countries affected by cyberattacks are the United States, Middle Eastern nations, Canada, Germany, and Japan. The 2024 Zscaler ThreatLabz report identifies the top five countries impacted by phishing attacks as the United States, the United Kingdom, India, Canada, and Germany [6].

The top five industries targeted by phishing scams, according to the IBM report [6], are Finance & Insurance, Manufacturing, Services, Technology, and Retail & Wholesale. Notably, the manufacturing sector is the most targeted by cybercriminals [5]. The CrowdStrike "Global Threat Report" from 2024 indicates that the most attacked sectors are Technology (23%), Telecommunications (15%), Financial (13%), Government (9%), Retail (9%), Manufacturing (8%), Healthcare (8%), Services (6%), Education (4%), and Media (4%) [7].

The most compromised accounts belong to Facebook, Instagram, Twitter, Snapchat, Twitch, Spotify, Gmail, Microsoft, Coinbase, and Amazon [8]. The three brands most frequently used in impersonation attacks have remained unchanged since 2019: Microsoft (57%), WeTransfer (17%), and DHL (6%) [9].

The education sector is particularly vulnerable, experiencing a 157% year-over-year increase in malware attacks between 2021 and 2022 [10]. Competition for global dominance—political, technological, economic, and military—has contributed to a significant rise in big-game hunting (BGH) adversaries, with 76% of such attacks potentially supported by nation-states [6]. Besides financial motives, these attacks target government agencies and any organization that holds sensitive data, including intellectual property, trade secrets, personal data, or medical records [6].

The cybersecurity landscape can change rapidly, making statistical data relevant for only a short period. New malware and variants can emerge daily; in 2023, antivirus testing identified over 100 million strains of malware and potentially unwanted applications (PUAs) [10]. However, reports from different times and sources can provide valuable insights for trend analysis.

2. Type of malware

Malware, or malicious software, comes in various forms, each designed to perform specific harmful activities. Common types of malwares include:

- Viruses:
 - *File Infectors*: Attach themselves to executable files and spread when the infected file is run.
 - *Macro Viruses*: Target applications like Microsoft Word or Excel, infecting documents or spreadsheets.
- Worms:
 - Standalone programs that replicate themselves to spread to other computers, often exploiting vulnerabilities in network security.
- Trojans:
 - *Backdoor Trojans*: Create a backdoor on the victim's system, allowing remote control by an attacker.
 - *Downloader Trojans*: Download and install other malicious software.
 - *Infostealers*: Steal sensitive information such as passwords or personal data.
- Ransomware:
 - Encrypts the victim's files or locks their system, demanding a ransom to restore access.
- Adware:
 - Displays unwanted advertisements on the victim's device, often within web browsers.
- Spyware:
 - Secretly monitors and collects information about a user's activities without their knowledge.
- Rootkits:
 - Modify the operating system to hide their presence, allowing continued privileged access to the system.
- Keyloggers:
 - Record keystrokes to capture sensitive information such as passwords and credit card numbers.
- Botnets:
 - Networks of infected computers (bots) controlled by an attacker to perform coordinated tasks like Distributed Denial of Service (DDoS) attacks.
- Phishing:
 - Uses fraudulent communication (often via email) to trick individuals into providing sensitive information.
- Smishing:
 - Uses SMS messages to trick individuals into providing sensitive information.
- Vishing:
 - Uses voice calls to trick individuals into providing sensitive information.
- Scareware or Hoaxes:

- Tricks users into believing their system is infected, prompting them to download or purchase unnecessary or harmful software.
- Fileless Malware:
 - Operates in memory, using legitimate system tools to perform malicious activities without leaving traces on the hard drive.

Each type of malware presents unique threats and requires specific strategies for detection and removal. These malware types can initiate various attacks and cause significant harm to infected systems.

It is estimated that there are around 1.2 billion malicious programs and Potentially Unwanted Applications (PUAs) [10]. In 2022, 81% of organizations experienced malware attacks, with 5.5 billion malware incidents worldwide—a 2% increase from the previous year [11].

Ransomware remains the most prevalent form of malware, accounting for nearly 70% of malware-related breaches and 24% of all breaches overall, with approximately 500 million ransomware attacks in 2022 [11]. It is estimated that someone falls victim to ransomware every 11 seconds [10]. Financial motivation drives 59% of ransomware attacks [11].

IoT (Internet of Things) malware variants and attacks are also on the rise. In 2022, there was an 87% year-over-year (YOY) increase in IoT malware, with a 252% YOY increase in the finance sector alone [11].

Estimates suggest that 65% of cyberattacks are linked to organized crime, 5% to nation-state or state-affiliated actors, around 15% to independent attackers (e.g., lone wolves), 3% to espionage, and around 1% to ideology or grudges [11]. Organized crime is responsible for 62% of ransomware incidents [11].

In 2022, Ukraine saw an 8,105% increase in malware attacks, with 5,835% related to ransomware [11].

The COVID-19 pandemic and related isolation measures significantly increased the demand for remote work, and by 2023, 62% of companies experienced security breaches related to remote work [12].

Cryptojacking—the unauthorized use of someone else's device to mine cryptocurrency, often through the use of XMRig mining malware—rose by 399% globally [13].

Malware-free attacks, which use low-observable characteristics (LOC), have increased in the past five years. In 2023, 75% of detected cyber activities involved malware-free attacks [6]. These types of attacks leave no footprint, making detection and removal particularly challenging.

Additionally, 84% of cybersecurity experts believe that supply-chain attacks, which exploit vulnerabilities in software dependencies, could become the biggest cyber threats for organizations [14]. Supply-chain attacks rose by 430% [14].

Fileless malware cyberattacks also surged by 75% in 2023 [6], making them one of the most difficult forms of attack to detect and mitigate, as they leave no visible trace on the system.

3. Social Engineering

"Social engineering refers to the methods attackers use to deceive victims into performing actions, typically opening a malicious webpage or running an unwanted file attachment" [15]. It is a form of psychological manipulation aimed at tricking victims into actions that may install harmful software, reveal sensitive information, or facilitate unauthorized access. These actions can lead to illegal activities such as unauthorized access to devices, theft of sensitive data or bank accounts, installation of harmful software, financial losses, or even espionage.

In North America, social engineering accounts for 56% of data breaches [14].

The consequences of such attacks vary depending on the malware involved, but they can be highly destructive. These include encrypting files, preventing access to computers, services, or equipment, stealing sensitive data, intellectual property, and causing severe business disruptions. For small businesses, the impact can be devastating: 60% of small businesses that experience a cyberattack are likely to go out of business within six months [16]. Many small businesses either cannot afford adequate cybersecurity investments or believe they won't be targeted (57%). However, small business breaches make up 43% of all data breaches [13].

3.1. Social Engineering Hacking

Social engineering attacks can be categorized based on the medium used to deliver malicious messages and deceive victims. These categories include:

1. Phishing – Distributed through contaminated email messages.
2. Smishing – Spread via SMS messages.
3. Vishing – Carried out through voice calls.
4. Pharming – Redirects users to fake websites that imitate legitimate ones to steal personal information, such as user IDs, passwords, or credit card information.

All these categories exploit human psychology and rely on manipulation techniques known as Social Engineering. It's estimated that 95% of cyber-attacks use some form of Social Engineering [10].

Social Engineering is a dominant method for delivering malware. In 2024, 84% of businesses and 83% of charities experienced some form of cybersecurity breach or attack [17]. Among these, phishing is the most prevalent form of attack, accounting for 51% of all Social Engineering-related incidents in 2023 [9]. According to a Statista report, 1,270,883 unique phishing sites were detected globally, with phishing attacks primarily targeting financial institutions. Around 18% of these attacks targeted web-based software and webmail services. Additionally, smishing affected 76% of global businesses, impacting 599 brands in October 2022 alone [19].

"In the first quarter of 2024, over 963,000 unique phishing sites were detected worldwide, marking a slight decrease from the previous quarter. The largest spike in phishing sites occurred between the second and third quarters of 2020, jumping from nearly 147,000 to approximately 572,000" [19].

The FBI has tracked significant growth in phishing, smishing, vishing, and pharming attacks. Reported cases rose from 25,344 in 2017 to 323,792 in 2021. In contrast, non-payment and non-delivery cases fell slightly, from 84,079 to 82,478 [12]. Personal data breaches have increased from 30,904 in 2017 to 51,629 in 2021, and identity theft has risen threefold in the same period. Extortion cases also rose from 14,938 in 2017 to 39,360 in 2021, though they dropped from 76,741 in 2020.

Globally, the cost of cybercrime has surged from \$3 trillion USD in 2015 to an estimated \$9.5 trillion USD in 2024. These costs include data damage, theft of intellectual property and personal data, lost productivity, post-attack disruption, legal costs, reputational harm, and more [18].

Ransomware is evolving, with hacking groups like Alphv, Hive, Conti, LockBit, 8Base, BlackBasta, and Play offering Ransomware-as-a-Service (RaaS) [18]. One of the most recent examples is the ransomware attack on MGM Resorts. While MGM reportedly suffered \$100 million USD in losses after refusing to pay the ransom, Caesar's Entertainment opted to pay \$15 million USD to prevent further costs and the disclosure of stolen data [18]. The hackers, known as Scattered Spider, may be a subgroup of the Alphv ransomware gang. They exploited MGM's systems, causing massive disruptions to services such as slot machines, ATMs, credit card payments, and online reservations [20].

Although MGM believes that the stolen customer data has not yet been used for identity theft, the possibility remains that it could be leveraged in future cyber-attacks. Experts are still investigating how the attack was carried out, with speculation that Scattered Spider used vishing, spoofed phone calls, and impersonation to obtain login credentials and deploy ransomware.

This example highlights the devastating effects of cyberattacks, disrupting critical services and causing significant financial damage. Social Engineering tactics are surprisingly effective, with the number of successful attacks proving how easily people can be deceived. Every day, individuals are subject to psychological manipulation, from marketing campaigns to political and financial messages across mass media, such as newspapers, TV, and digital platforms.

The growing accessibility of phishing kits also makes it easier for attackers with minimal technical knowledge to launch phishing attacks, contributing to the prevalence of Social Engineering-based cybercrime.

4. Artificial Intelligence

The rise of Artificial Intelligence (AI) and Generative AI (GenAI) has sparked considerable debate in cybersecurity, as these technologies are anticipated to both revolutionize the field and introduce new

challenges. While AI and GenAI hold promise for advancing cybersecurity, there are concerns about their potential to increase vulnerabilities and enhance the capabilities of cybercriminals.

A 2024 Verizon report suggests that AI and GenAI have not yet made groundbreaking advances in pretexting or phishing—techniques that are already highly effective. Deepfake technology, however, is considered an exception, with numerous frauds already committed using this method. According to a survey, 46% of senior security professionals believe that GenAI will heighten vulnerabilities, illustrating the uncertainty surrounding its future impact.

One of the most concerning developments is the increasing sophistication of deepfake technology. GenAI can now generate hyper-realistic videos and images that not only deceive people but can also trick security software. This presents a major challenge in the realm of phishing and vishing (voice phishing) attacks, as criminals can exploit GenAI to create convincing fake media that facilitates social engineering schemes. AI-generated deepfakes have the potential to bypass biometric security systems, such as fingerprints, facial recognition, and voice authentication.

For example, CAPTCHA—the traditional test to distinguish between humans and bots—can be defeated through AI-powered image analysis. Similarly, voice cloning, powered by GenAI, has been used to impersonate individuals, deceiving not only common users but also subordinates or even executives into divulging sensitive information. The FBI has reported cases where cybercriminals used synthetic media and platforms like WhatsApp to impersonate company executives for financial gain. One particular incident involved an unknown actor impersonating an executive over a poor-quality WhatsApp call, suggesting a switch to a video conference. Though the target became suspicious and terminated the call, the incident highlights the growing use of AI in vishing attacks.

In addition to deepfake-based attacks, hackers have developed malicious AI-powered tools like WormGPT and FraudGPT. These tools mimic legitimate AI chatbots, like ChatGPT, but are designed for illicit purposes, enabling cybercriminals to automate and scale attacks. The introduction of customizable chatbots that can act as personal assistants is expected to lead to a surge in AI users, potentially escalating the current threat landscape.

Voice cloning, in particular, has emerged as a terrifying application of GenAI in cybercrime. Criminals have increasingly used cloned voices in vishing schemes to simulate family member kidnappings or other emergencies, extorting money from victims. Alarming, voice cloning services are widely available—basic versions can be found for free, while more sophisticated tools range from \$20 to several hundred dollars per month.

The combination of AI, deepfakes, and voice cloning is reshaping the cybersecurity landscape. As GenAI evolves, both defenders and attackers will face new challenges. While AI presents opportunities to strengthen cybersecurity systems, it simultaneously arms cybercriminals with.

5. Conclusion

Cybercrime is escalating at an alarming rate, posing significant risks to individuals, businesses, and critical infrastructure. Fraud, identity theft, ransomware, data breaches, and misinformation have become unavoidable parts of daily life. As digital technologies expand, so too does the vulnerability to cyberattacks. From sophisticated zero-day exploits to Social Engineering schemes, cybercriminals now have access to powerful tools that require minimal technical expertise, including DIY phishing kits. The insider threat, whether intentional or accidental, continues to be a major concern, as a single click can lead to devastating consequences for organizations and institutions.

The increasing dependence on digital platforms for managing critical resources, such as power, water, and communication systems, has made these infrastructures susceptible to remote cyber sabotage. While traditional physical sabotage remains a threat, the ability to launch digital attacks from across the globe has added a new dimension to warfare. The sabotage of the Nord Stream pipeline in 2022 serves as a stark reminder of the vulnerabilities that exist in both physical and digital realms.

The rise of AI, especially Generative AI, is transforming the cybersecurity landscape. While AI has the potential to enhance security through improved analytics, it also introduces new risks. Deepfake technology, voice cloning, and the creation of sophisticated phishing attacks are just a few examples of how AI can be exploited. Moreover, AI-driven tools can enable even those with limited technical skills to launch complex cyberattacks, increasing the overall threat landscape.

Supply-chain attacks represent another growing concern, as software dependencies and third-party components create opportunities for malicious actors to inject malware or open backdoors in critical systems. In such cases, the best-case scenario is often financial loss; the worst-case scenario could be catastrophic.

The shifting international landscape, with new power centres and unclear spheres of influence, has heightened awareness of cybersecurity's importance. The absence of a unified set of international rules governing AI and cybersecurity exacerbates the uncertainty. As history shows, humanity's arrogance in believing it can foresee and mitigate all risks is often misguided. We are frequently caught off-guard by emerging threats, especially those that bring widespread harm.

In a world with rising geopolitical tensions and unclear global governance, it is difficult to predict the future. However, it is imperative that we work towards establishing clear international norms around AI and cybersecurity. This is not just for the protection of individual nations but for the safety of all humankind and the preservation of the natural world.

References:

- [1] UpGuard, "Biggest Data Breaches in USA History (Updated 2024)", <https://www.upguard.com/blog/biggest-data-breaches-us#toc-0>.
- [2] Computerworld, "Maersk-selskab ramt af ransomware-angreb: Store datamængder svæver på det mørke web", <https://www.computerworld.dk/art/263111/maersk-selskab-ramt-af-ransomware-angreb-store-datamaengder-svaever-paa-det-moerke-web>.
- [3] Wikipedia, "Petya (malware family)", [https://en.wikipedia.org/wiki/Petya_\(malware_family\)](https://en.wikipedia.org/wiki/Petya_(malware_family)).
- [4] IBM, "Cost of a Data Breach Report (2023)", available online at <https://www.ibm.com/reports/data-breach#:~:text=The%20global%20average%20cost%20of,15%25%20increase%20over%203%20years.&text=51%25%20of%20organizations%20are%20planning,threat%20detection%20and%20response%20tools>.
- [5] Zscaler, "Zscaler ThreatLabz 2024 Phishing Report", available at <https://www.zscaler.com/campaign/threatlabz-phishing-report>.
- [6] CrowdStrike, "Global Threat Report (2024)", available online at https://go.crowdstrike.com/global-threat-report-2024.html?utm_campaign=brand&utm_content=crwd-brand-eur-nord-en-psp-x-trl-x-tct-x_x_x_reports-x&utm_medium=sem&utm_source=goog&utm_term=crowdstrike%202024%20global%20threat%20report&cc_cmp=20059710141&cc_plac=&gad_source=1.
- [7] VPN Central, "Which Accounts Are Americans Losing to Hackers the most? (2023)", <https://vpncentral.com/most-hacked-accounts-study/>.
- [8] Aricioğlu, Beyza, "Social Engineering Statistics to Know in 2024", RESMO, available on Internet at <https://www.resmo.com/blog/social-engineering-statistics>.
- [9] Barracuda (2022) "Spear-phishing report: Social Engineering and growing complexity of attacks", available online at <https://blog.barracuda.com/2022/03/16/spear-phishing-report-social-engineering-and-growing-complexity-of-attacks>.
- [10] Petrosyan Ani (2024), "Number of global phishing Q3 2023– Q1 2024", Statista, available online at <https://www.statista.com/statistics/266155/number-of-phishing-domain-names-worldwide/>.
- [11] Smith Gary (2024), "65 Malware Statistics for 2024", StationX, available online at <https://www.stationx.net/malware-statistics>.
- [12] FBI (2021), "Federal Bureau of Investigation Internet Crime Report 2021", Internet Complaint Center, US, available online at https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf.
- [13] Verizon (2024), "Data Breach Investigation Report", Verizon, available online at <https://www.verizon.com/business/resources/reports/2024-dbir-data-breach-investigations-report.pdf>.
- [14] Lenaerts-Bergmans Bart (2023), "What is a Supply-Chain Attack?", CrowdStrike, available online at <https://www.crowdstrike.com/cybersecurity-101/cyberattacks/supply-chain-attacks/>.
- [15] Sophos LTD (2017), "Threatsaurus: The A – Z of computer and data security treats", Oxford, UK.

- [16] Galvin Jo (2018), “60 Percent of Small Businesses Fold Within 6 Months of a Cyber Attack. Here's How to Protect Yourself” Vistage, available online at <https://www.inc.com/joe-galvin/60-percent-of-small-businesses-fold-within-6-months-of-a-cyber-attack-heres-how-to-protect-yourself.html>.
- [17] GOV.UK (2024), “Cyber security breaches survey 2024”, Official statistics, GOV.UK, available online at <https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2024/cyber-security-breaches-survey-2024#chapter-7-conclusions>.
- [18] Cybersecurity Ventures (2024), “Cybercrime To Cost The World \$9.5 Trillion USD Annually In 2024”, Cybersecurity Ventures , Essentire, available online at https://www.esentire.com/web-native-pages/cybercrime-to-cost-the-world-9-5-trillion-usd-annually-in-2024?utm_medium=email&utm_source=pardot&utm_campaign=autoresponder.
- [19] Statista (2024), “Phishing – Statistics & Facts”, Statista, available online at <https://www.statista.com/topics/8385/phishing/#topicOverview>.
- [20] Gomez Fatima (2024), “2023 Cyber Attack & Breach on the MGM Resort Explained”, INSZONE, available online at <https://inszoneinsurance.com/blog/cyberattack-mgm-resort-explained>.
- [21] Sosafe (2024) “The Cybercrime trends to watch in 2024”, Sosafe, available online at <https://sosafe-awareness.com/resources/reports/cybercrime-trends/>.

Investigating the Impact of Social Networks on Viral Marketing

Magdalena Ninkov¹, Snežana Jokić², Đorđe Alavuk¹, Nikola Jančev¹, Marjana Pardanjac²

¹ Higher Business School of Vocational Studies, Novi Sad, Serbia

² University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

magdalenaninkov@vps.ns.ac.rs, snezana.jokic@tfzr.rs, marjana.pardanjac@tfzr.rs,
djordjealavuk@yahoo.com, jancevn@gmail.com

Abstract:

Social networks are essential in viral marketing because they quickly share information among many users. Sharing content becomes viral and reaches a broad audience without significant marketing costs. In addition to the virality of content, social networks enable communication between brands and users, strengthening trust and loyalty. This paper demonstrates the impact of social networks on viral marketing. The Kruskal-Wallis test was used to determine whether there were statistically significant differences between the groups of respondents to analyze the results. Where it was determined that there were statistically significant differences, a Post Hoc test was performed to determine between which groups the difference occurred. Based on the results obtained, recommendations are given for social network purposes in viral marketing.

Keywords:

Marketing, viral marketing, social networks.

1. Introduction

In the modern digital age, marketing is one of the most critical factors for the success of any organization. Marketing is the link between the company and the consumer, and its importance is reflected in the ability to increase brand awareness, retain existing and attract new customers, and thus enable business growth. Viral marketing stands out as an essential segment of marketing. Viral marketing is a powerful strategy in the digital world - in general. It is defined as a subtype of marketing that causes interest and potential sales of products, i.e. services, through messages that spread like a "virus" on social networks. A key feature of viral marketing is its ability to spread organically through users sharing content with their friends, thus achieving a broad reach, all at a relatively low cost. Social networks are platforms used to connect users, and they have become a daily routine for people worldwide. In this paper, the emphasis will be on social networks and how they can contribute to viral marketing. The integration of viral marketing and social networks allows brands to reach a wide range of audiences through the organic sharing of content among users, which could result in exponential growth in visibility and engagement.

1.1. The phenomenon and evolution of social networks

Social networks have become a global communication phenomenon because man is inherently a social being, and social communication interests and attracts him. People use social networks to get closer and connect with other people, exchanging information and expressing opinions. As a common feature of all social networks, universal means of communication and techniques are mentioned, but also a common interest that holds all people together and therefore gives enormous popularity to all social networks today (Grbavac & Grbavac, 2014). The BBC is the first social network most similar to today's social networks (Bulletin et al.). It appeared around 1978 and allowed users to have some essential communication through messages. The first popular website, "Six Degrees", was launched around 1997.

It is based on the "Theory of Six Steps", which aims to "reach" all the people in this world through approximately six steps. Users of this social network can create user profiles, search for them, and communicate with friends. This site was discontinued around 2000 because it has not proven viable (Boyd & Ellison, 2008). In the period from 1997 to 2001, several more social networks were developed - such as AsianAvenue" for the Asian American population, BlackPlanet (for the African American population) and MiGente (for the Latino population). In that period, LiveJournal (an online diary that could be read and commented on) and the social network "Friendster" appeared, which did not survive in the long run due to the large number of fake profiles that appeared at that time (Horvatin, 2014).

1.2. Types of social networks

At first glance, it seems that social networks are easy to define; however, from the attached, it can be seen that this is not quite so (Whiting & Williams, 2013). Many scientific works (both domestic and foreign) recently deal with the concept of social networks and their features and definitions. (Boyd & Ellison, 2008). Due to user activities and the purpose that defines a social network, most social networks are placed in several general categories. (Ružić, Turkalj, & Biloš, 2014):

- **Community social networks** – this type of social network includes all web pages. Its purpose is to make it easier for users to expand their social networks through existing contacts. Community social networks represent the global Internet community, and an example of such a network is: "MySpace". Here, smaller groups are formed based on common interests. Other social networks are focused on specific demographic groups (e.g. "BlackPlanet and MiGente"). This group is the most numerous social network, including Facebook, Twitter and Google +.
- **Social networks of professionals** - they represent a subgroup of the previous group. One or more specificities connect the members of these communities. An example of this group social network is LinkedIn (a social network for business professionals).
- **Networks for the exchange of multimedia content** - the primary "activity" of this group social network is the publication of images, videos and audio recordings. The most popular network for sharing audio content is YouTube, and networks for sharing visual content are Instagram, Flickr and others.
- **Music social networks** - could be included in the previous group of networks, but according to some experts, they stand out as a separate category. The role of these social networks is to share music with friends. All users can search and listen to their favorite artists as their songs. It is possible to rank and comment on playlists, and based on ratings, the network receives information about user preferences. An example of such a social network is YouTube.
- **Social networks for bookmarking** - Bookmarking is the practice of saving web links in web browsers. In some browsers, these tags are called bookmarks; in others, they are called favorites. The point is to save the same web link in the list on the social network. When users save a link, it is possible to group them and mark them according to keywords ("tags"). An example is Pinterest.
- **Social networks of bloggers** - a place for a standard profile; each user runs and writes his blog and edits it with a topic of his choice. It is possible to form smaller communities with other bloggers.

Social networks have changed the world and become part of people's daily interactions. Currently, there are many modern networks. Some are popular, and some are more popular. Their popularity depends on people's adaptation to new technologies to retain existing users and attract new ones.

2. Research methods

Relevant literature and empirical research from thematic journals were used to write this paper. The collected data were examined using qualitative methods. The paper aims to investigate which social networks respondents use most often and how they can be used for viral marketing. The research was carried out using a voluntary and anonymous survey. The questionnaire was structured specifically for this research and based on the literature study that enabled the definition of the research variables. The survey questionnaire is divided into two parts. The first part includes the socio-demographic characteristics of the respondents, and the second part involves questions related to social network

variables and viral marketing. Multiple-choice questions and a Likert scale are used in the questionnaire. The questionnaire was compiled in Google Forms and distributed electronically. Sampling was done using random selection, and the sample size was 231 respondents (see Table 1) from the Vojvodina area. The research method was used to obtain the necessary data on the socio-demographic characteristics of the respondents/students, the most often used social networks among this population, and the habits and attitudes of the population regarding this topic. The collected research data were statistically processed in SPSS and presented in tabular form. Since the obtained data are not normally distributed, the Kruskal-Wallis H and Mann-Whitney tests were used for statistical processing. These are non-parametric tests used to determine whether there are statistically significant differences between two, i.e. multiple subgroups of independent variables.

3. Findings and discussion

The general characteristics of the respondents are shown in Table 1.

Table 1
General Characteristics of the Respondents

		Frequency	Per cent	Valid Percent	Cumulative Percent
Sex	Male	75	32.5	32.5	32.5
	Female	156	67.5	67.5	100
	Total	231	100	100	
Age of the Respondents	Less than 18	6	2.6	2.6	2.6
	18 – 25	119	51.5	51.5	54.1
	26 – 35	26	11.3	11.3	65.4
	36 – 45	27	11.7	11.7	77.1
	46 – 55	46	19.9	19.9	97
	Over 55	7	3.0	3.0	100
	Total	231	100	100	
Employment status	Pupil	22	9.5	9.5	9.5
	Student	92	39.8	39.8	49.3
	Unemployed	9	3.9	3.9	53.2
	Employed	104	45.0	45	98.2
	Retired	4	1.7	1.8	100
	Total	231	100	100	

Based on the previously processed data in the table, it can be concluded that the answers to the questions regarding *the gender of the respondents* were mostly given by female respondents (156 respondents, i.e. 67.50%), compared to male respondents (75 respondents, i.e. 32, 50%). Although the deviation between the sexes is not very small, this structure of responses was not crucial for further investigation. When it comes to the *age of the respondents*, based on the statistically obtained results, it is concluded that the questionnaire was mostly filled out by the young population aged 18-25 years, i.e. 119 respondents, which is 51.50% of the total number of respondents. Then, there are respondents between the ages of 46 and 55, which is expressed as a percentage of 19.90%. Respondents between the ages of 36 and 45 take third place. There are 11.70% of these respondents. Only one less respondent is in the group of respondents aged between 26 and 35. Therefore, there are 26 of these respondents, which is 11.30% of the total respondents. The penultimate place is occupied by respondents over 55 years, who are 3% of the total respondents. In the last place, there is a group of respondents less than 18 years of age, i.e. of these respondents is 2.60%. It can be concluded that more than half of the respondents (51.50%) are aged 18-25. This result is entirely acceptable; given that this is the most represented category on social networks. About the work status of the respondents, the most significant number of respondents have the status of a student - 92 of them, i.e. 39.80%. Students and employees shared the second place, with 22 each (9.50%). The penultimate place is occupied by unemployed

respondents (9 respondents, that is, 3.90%). In the last place are pensioners, of whom there are only four, i.e. 1.70%).

Table 2

Descriptive statistics of social networks

	Frequency	Per cent	Valid Percent	N	Mean	Std. Deviation
Instagram	192	79.7	83.1	231	.83	.375
YouTube	146	60.6	63.2	231	.63	.483
Facebook	106	44.0	45.9	231	.46	.508
TikTok	106	44.0	45.9	231	.46	.499
Snapchat	57	23.7	24.7	231	.25	.432
Pinterestt	38	15.8	16.5	231	.16	.372
LinkedIn	11	4.6	4.8	231	.05	.213
Twitter	9	3.7	3.9	231	.04	.194

Table 2 presents descriptive statistics for the factors of the social networks used by the respondents. Social networks are ranked based on the mean value (Mean). Based on the responder's answers, we can conclude that they do not use all social networks equally. Collected data were investigated and ranked by use. Instagram is the most used network, with 83.1% of respondents. Then, YouTube is in second place and is used by 63.2% of respondents. The third place is the Facebook network, with 45.9% of respondents. TikTok took fourth place with 45.9% of respondents. In fifth place is Snapchat, which is used by 24.7% of respondents. The Pinterest network takes the penultimate place with 16.5% of respondents. LinkedIn is used by 4.8% of respondents. In last place is Twitter, which had 3.9% of respondents. The standard deviation of the Twitter social network is .194, which indicates the slightest difference between respondents' answers; the reason for this may be the small number of respondents who use this network. The most significant difference between respondents' answers is observed in the social network Facebook (Std. Dev = .508).

3.1. Results by Gender

Evaluating the obtained data and researching whether there is a statistically significant difference between gender preferences with the social networks they use, the Mann-Whitney Test was determined, and the results are as follows:

Table 3

Test Statistics ^a

	Pinterest
Mann-Whitney U	4540.500
Wilcoxon W	7390.500
Z	-4.288
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: Sex

Table 4

Ranks

	Sex	N	Mean Rank	Sum of Ranks
Pinterest	Male	75	98.54	7390.50
	Female	156	124.39	19405.50
	Total	231		

The Mann-Whitney test results indicate a statistically significant difference between the sexes regarding the use of the Pinterest network in favor of the female population ($U=4540,500$, $N_1=75$, $N_2=156$, $p=0.000$). We notice that Female (Mean Rank) is higher than Male (Mean Rank). Based on the results, it is concluded that the Pinterest social network is convenient for addressing the female population and for viral marketing.

3.2. Results by age of the respondents

Table 5

Test Statistics ^{a,b}

	Facebook	Instagram	TikTok	YouTube	Snapchat	Twitter	Pinterest	LinkedIn
Chi-Square	36.602	17.212	35.606	10.066	36.159	4.302	5.859	16.207
df	5	5	5	5	5	5	5	5
Asymp. Sig.	.000	.004	.000	.073	.000	.507	.320	.006

a. Kruskal Wallis Test

b. Grouping Variable: Year

The Kruskal-Wallis test showed that there is a statistically significant difference between the age of the respondents and the use:

1. **Facebook networks** $H(5)=36.602$, $p<0.000$, $\eta^2=5.824$. The results show that the mean rank changes depending on the respondents' age; it is lowest for age group 1 and highest for age group 3.
2. **Instagram networks** $H(5)=17.212$, $p<0.004$, $\eta^2=1.288$. The results show that the mean rank changes depending on the respondents' age; it is lowest for age group 1 and highest for age group 3.
3. **TikTok networks** $H(5)=35.606$, $p<0.000$, $\eta^2=5.512$. The results show that the Mean Rank changes depending on the respondents' age; it is lowest for age group 4 and highest for age group 1.
4. **Snapchat networks** $H(5)=36.159$, $p<0.000$, $\eta^2=5.684$. The results show that the mean rank changes depending on the respondents' age; it is inversely proportional—the higher the rank, the lower the age, and vice versa.
5. **LinkedIn network** $H(5)=16.207$, $p<0.006$, $\eta^2=1.142$. The results show that the mean rank changes depending on the respondents' age; it is lowest for age group 1 and highest for age group 3.

After establishing statistically significant differences, Post Hoc testing with the Mann-Whitney test was performed to accurately identify different subgroups from one another (age group 1 - (less than 18), age group 2 - (18-25), age group 3 - (26-35), age group 4 - (36- 45), age group 5 - (46-55), age group 6 -over 55).

Table 6

Test Statistics Facebook

	1-3	1-4	1-5	1-6	2-3	2-4	2-5	2-6
Mann-Whitney U	21.000	27.000	51.000	6.000	881.000	1017.000	1831.000	244.000
Wilcoxon W	42.000	48.000	72.000	27.000	8021.000	8157.000	8971.000	7384.000
Z	-3.234	-2.921	-2.896	-2.535	-4.091	-3.561	-3.894	-2.275
Asymp. Sig. (2-tailed)	.001	.003	.004	.011	.000	.000	.000	.023
Exact Sig. [2*(1-tailed Sig.)]	.004 ^b	.010 ^b	.010 ^b	.035 ^b				

By post hoc testing of different age groups' use of the social network Facebook, it was determined that there is a difference between the following age groups: 1 and 3, 1 and 4, 1 and 5, 1 and 6, 2 and 3, 2 and 4, 2 and 5, and 2 and 6. The results indicate differences in the use of the social network Facebook between respondents younger and older than 26 years, which should be considered when using this social network for viral marketing purposes.

Table 7

Test Statistics ^{a,b} Instagram

	2-5	2-6	3-5	3-6
Mann-Whitney U	2143.500	283.500	449.000	59.000
Wilcoxon W	3224.500	311.500	1530.000	87.000
Z	-3.317	-2.456	-2.375	-2.268
Asymp. Sig. (2-tailed)	.001	.014	.018	.023
Exact Sig. [2*(1-tailed Sig.)]				.169 ^b

By post hoc testing of different age groups' use of the social network Instagram, it was determined that there is a difference between the following age groups: 2 and 5, 2 and 6, 3 and 5, 3 and 6. The results indicate the differences in the use of the Instagram social network between respondents aged 18 to 35 and those over 46.

Table 8Test Statistics ^{a,b} TikTok

	1-3	1-4	1-5	1-6	2-3	2-4	2-5
Mann-Whitney U	27.000	15.000	33.000	9.000	1146.500	932.000	1735.500
Wilcoxon W	378.000	393.000	1114.000	37.000	1497.500	1310.000	2816.500
Z	-2.847	-3.771	-3.701	-2.138	-2.400	-3.932	-4.202
Asymp. Sig. (2-tailed)	.004	.000	.000	.033	.016	.000	.000
Exact Sig. [2*(1-tailed Sig.)]	.012 ^b	.001 ^b	.001 ^b	.101 ^b			

By post hoc testing of different age groups in the use of the social network TikTok, it was determined that there is a difference between the following age groups: 1 and 3, 1 and 4, 1 and 5, 1 and 6, 2 and 3, 2 and 4, 2 and 5. The results indicate that there are differences in the use of the social network TikTok between respondents under the age of 18 and respondents over the age of 26, as well as between respondents aged 18 to 25 and respondents aged 26 to 55.

Table 9Test Statistics ^{a,b} Snapchat

	1-3	1-4	1-5	1-6	2-4	2-5	2-6	3-5
Mann-Whitney U	41.000	33.000	49.000	7.000	1118.000	1761.500	259.000	496.000
Wilcoxon W	392.000	411.000	1130.000	35.000	1496.000	2842.500	287.000	1577.000
Z	-2.293	-3.352	-4.991	-2.494	-3.043	-4.564	-2.021	-2.498
Asymp. Sig. (2-tailed)	.022	.001	.000	.013	.002	.000	.043	.012
Exact Sig. [2*(1-tailed Sig.)]	.078 ^b	.024 ^b	.009 ^b	.051 ^b				

Post hoc testing of different age groups in the use of the Snapchat social network found that there was a difference between the following age groups: 1 and 3, 1 and 4, 1 and 5, 1 and 6, 2 and 4, 2 and 5, 2 and 6, 3 and 5. The results indicate that there are differences in the use of the social network Snapchat between respondents younger than 18 years and respondents older than 26 years, then respondents aged 18 to 25 years and respondents 36 to 55 years old, as well as respondents aged from 26 to 35 years old with respondents from 46 to 55 years old.

Table 10Test Statistics ^{a,b} LinkedIn

	2-3	2-6	3-5
Mann-Whitney U	1275.500	364.000	509.000
Wilcoxon W	8415.500	7504.000	1590.000
Z	-3.769	-2.117	-2.033
Asymp. Sig. (2-tailed)	.000	.034	.042

Post-hoc testing of different age groups regarding the use of the social network LinkedIn determined a difference between the following age groups: 2 and 3, 2 and 6, 3 and 5. The results indicate that there are differences in the use of the social network LinkedIn between respondents from 18 to 25 years old, respondents from 26 to 35 years old and respondents older than 55 years old, and respondents from 26 to 35 years old with respondents from 46 to 55 years old.

Based on the results, the most statistical differences were observed among the respondents in the age groups 18 to 25, 36 to 45, 18 to 25, and 46 to 55; 4 statistically significant differences were observed. With social networks such as Facebook and Snapchat, seven statistically significant differences were observed between the groups of respondents.

3.3. Results by work status

Table 11

Test Statistics^{a,b}

	Facebook	Instagram	TikTok	YouTube	Snapchat	Twitter	Pinterest	LinkedIn
Chi-Square	43.336	9.689	35.906	4.785	29.775	2.482	4.865	11.326
df	4	4	4	4	4	4	4	4
Asymp. Sig.	.000	.046	.000	.310	.000	.648	.301	.023

a. Kruskal Wallis Test

b. Grouping Variable: Employment

The Kruskal-Wallis test showed that there is a statistically significant difference between the respondent's work status and the use of the following social networks:

1. **Facebook** $H(4)=43.336$, $p<0.000$, $\eta^2=8.165$. The results show that the mean rank changes depending on the respondent's work status; it is the lowest for group 2 and the highest for group 4.
2. **Instagram** $H(4)=9.689$, $p<0.000$, $\eta^2=0.408$. The results show that the Mean Rank changes depending on the respondent's work status. It is lowest for group 2 and highest for group 3.
3. **TikTok** $H(4)=35.906$, $p<0.000$, $\eta^2=5.605$. The results show that the Mean Rank changes depending on the respondent's work status. It is lowest for group 1 and highest for group 2.
4. **Snapchat** $H(4)=29.775$, $p<0.000$, $\eta^2=3.854$. The results show that the Mean Rank changes depending on the respondent's work status. It is lowest for group 5 and highest for group 2.
5. **LinkedIn** $H(4)=43.336$, $p<0.000$, $\eta^2=8.165$. The results show that the mean rank changes depending on the respondents' work status; it is the lowest for groups 1 and 2 and the highest for group 5.

After establishing statistically significant differences, Post Hoc testing with the Mann-Whitney test was performed to accurately identify different subgroups from one another (1 - unemployed, 2 - student, 3 - student, 4 - employed, 5 - retired) in the respondent's work status group.

Table 12

Test Statistics^{a,b} TikTok

	1-2	1-3	2-4	2-5	3-4
Mann-Whitney U	40.000	254.000	527.000	19.000	3206.000
Wilcoxon W	85.000	299.000	5987.000	29.000	8666.000
Z	-3.097	-2.227	-4.733	-2.311	-4.638
Asymp. Sig. (2-tailed)	.002	.026	.000	.021	.000
Exact Sig. [2*(1-tailed Sig.)]	.009 ^b			.081 ^b	

By post hoc testing of different subgroups of respondents by work status in the use of TikTok, it was determined that there is a difference between the following subgroups: 1 and 2, 1 and 3, 2 and 4, 2 and 5, 3 and 4. The results indicate the existence of different uses of the social network TikTok between respondents of employment subgroups, which should be taken into account when using this network for viral marketing purposes.

Table 13

Test Statistics^{a,b} Snapchat

	1-2	2-3	2-4	2-5	3-4
Mann-Whitney U	51.500	755.000	600.000	18.000	3724.000
Wilcoxon W	96.500	5033.000	6060.000	28.000	9184.000
Z	-2.397	-2.188	-5.061	-2.132	-3.731
Asymp. Sig. (2-tailed)	.017	.029	.000	.033	.000
Exact Sig. [2*(1-tailed Sig.)]	.037 ^b			.069 ^b	

Post hoc testing of different subgroups of respondents by work status in the use of Snapchat determined a difference between subgroups 1 and 2, 2 and 3, 2 and 4, 2 and 5, 3 and 4. The results indicate the existence of different uses of the social network Snapchat between respondents of employment subgroups, which should be taken into account when using this network for viral marketing purposes.

Table 14Test Statistics ^{a,b} Facebook, Instagram and LinkedIn

	Facebook			Instagram		LinkedIn		
	1-4	2-4	3-4	2-3	3-4	2-5	3-4	3-5
Mann-Whitney U	247.000	558.000	2811.000	813.000	4090.000	33.000	4422.000	140.000
Wilcoxon W	292.000	811.000	7089.000	1066.000	9550.000	286.000	8700.000	4418.000
Z	-2.823	-4.418	-5.734	-2.595	-2.808	-2.345	-2.397	-3.261
Asymp. Sig. (2-tailed)	.005	.000	.000	.009	.005	.019	.017	.001
Exact Sig. [2*(1-tailed Sig.)]						.471 ^b		.440 ^b

a. Grouping Variable: Employment

b. Not corrected for ties.

By post hoc testing of different subgroups of respondents by work status in the use of the social network Facebook, it was determined that there is a difference between the following groups: 1 and 4, 2 and 4, 3 and 4. The results indicate different uses of the social network Facebook between the respondent's different employment subgroups, which should be taken care of when using this network for viral marketing purposes.

Post hoc testing of different subgroups of respondents by work status group in the uses of the Instagram social network show a difference between the following subgroups: 2 and 3, 3 and 4. The results indicate differences in the use of the Instagram social network between respondents of different employment groups, which should be considered when using this network for viral marketing purposes.

Post hoc testing of different subgroups of respondents by work status group in the use of the LinkedIn social network show a difference between the following subgroups: 2 and 5, 3 and 4, 3 and 5. The results indicate differences in the use of the LinkedIn social network between respondents of different employment groups, which should be taken care of when using this network for viral marketing purposes.

The results show the most statistical differences between students and employees (5) subgroups. For the social networks TikTok and Snapchat, five statistically significant differences were observed between the subgroups of respondents.

4. Discussion

The research findings illustrate the significant role that social networks play in viral marketing and how various demographic factors, such as age, gender, and employment status, influence social network usage. The analysis, conducted using non-parametric tests (Kruskal-Wallis and Mann-Whitney), reveals that not all social networks are equally effective in reaching all demographic groups. For instance, 83.1% of respondents use Instagram as the most popular platform, followed by YouTube with 63.2%. These networks offer great potential for viral marketing, especially considering their widespread use among younger audiences.

The results are consistent with other studies that have found Instagram and YouTube to be highly influential in viral marketing, given their massive user base and ease of content sharing (Baabcha et al., 2022). These platforms facilitate the fast spread of information, particularly visual content, which often performs better in engaging users and encouraging shares. Additionally, viral marketing benefits from platforms embedded with strong social ties, which enhance trust and the likelihood of content dissemination (Borges-Tiago et al., 2018).

The results also indicate that Pinterest appeals more to women, making it an ideal platform for targeting female audiences through viral marketing. This finding aligns with previous research showing

that content tailored to specific demographics performs better in engagement and sharing (Akyol, 2013). Similarly, TikTok and Snapchat have a strong presence among younger users, particularly those aged 18-25, such as students and pupils. This suggests that these platforms are well-suited for engaging younger audiences in viral marketing campaigns, as supported by findings from other studies focusing on the effectiveness of firm-generated content across different platforms (Supartiwi et al., 2023).

On the other hand, Facebook is more frequently used by older groups (ages 26-55) and employed individuals, making it a strong contender for targeting mature, professional demographics. This is consistent with prior studies that emphasise the role of Facebook in reaching older age groups, who tend to use it more frequently for professional or personal communication (Ahmed & Abdelkader, 2018). LinkedIn, a network for professionals, is best utilised to address unemployed or career-oriented individuals. Research confirms LinkedIn's effectiveness in engaging individuals looking for career-related content, further demonstrating the importance of selecting the right platform for the right audience (Baabcha et al., 2022).

The key takeaway from this research is that viral marketing strategies must be tailored to specific social networks based on the audience's demographics and usage patterns. This differentiation enables more effective communication, higher engagement rates, and a better likelihood of a viral campaign. As Borges-Tiago et al. (2018) noted, content that aligns with the emotional tone and arousal level preferred by the platform's demographic will significantly enhance its virality.

5. Conclusion

The study confirms that social networks are vital for viral marketing, providing platforms where users organically share content, leading to rapid and widespread visibility. The research highlights significant differences in social network usage across various demographic groups, which must be considered when designing viral marketing campaigns. The main conclusions include:

1. Instagram and YouTube are the most widely used networks, making them essential platforms for viral marketing campaigns targeting broad, young demographics.
2. Pinterest is particularly effective for campaigns aimed at the female population.
3. Given their popularity among younger audiences, TikTok and Snapchat are ideal for engaging them, particularly students and teenagers.
4. Facebook is more suitable for older demographics, especially individuals aged 26-55, and for targeting employed individuals.
5. LinkedIn is best utilised to reach unemployed or professional individuals, especially those seeking career-related content.

The findings suggest that marketers should carefully select their social media platforms based on the target audience's demographic characteristics to maximise the effectiveness of viral marketing campaigns. Future research should explore more extensive and diverse samples to validate these insights further and delve into the specific content types that drive viral sharing on different platforms.

References

- [1] Ahmed, H. M. S., & Abdelkader, A. A. (2018). The Mediating Role of Repositioning on the Relationship Between Electronic Integrated Communication Marketing (E-IMC) and Loyalty. Faculty of Commerce, Kafrelsheikh University, Egypt.
- [2] Akyol, Ş. (2013). Social Media and Marketing: Viral Marketing. *Academic Journal of Interdisciplinary Studies*, 2(8), 586-593. DOI:10.5901/ajis.2013.v2n8p586.
- [3] Baabcha, H., Laifa, M., & Akhrouf, S. (2022). Social Influence Analysis in Online Social Networks for Viral Marketing: A Survey. In *International Conference on Managing Business Through Web Analytics*, (pp. 165-176). Springer, Cham. https://doi.org/10.1007/978-3-031-06971-0_11
- [4] Baabcha, H., Laifa, M., Akhrouf, S. (2022). Social Influence Analysis in Online Social Networks for Viral Marketing: A Survey. In: Sedkaoui, S., Khelfaoui, M., Benaichouba, R., Mohammed Belkebir, K. (eds) *International Conference on Managing Business Through Web Analytics*. Springer, Cham. https://doi.org/10.1007/978-3-031-06971-0_11

- [5] Borges-Tiago, M. T., Tiago, F., & Cosme, C. (2018). Exploring users' motivations to participate in viral communication on social media. *Journal of Business Research*, 96, 264-271. <https://doi.org/10.1016/j.jbusres.2018.11.011>
- [6] Boyd, D., & Ellison, N. (2008). *Social Network Sites: Definition, History*, *Journal of Computer-Mediated Communication*, 210-230.
- [7] Grbavac, J., & Grbavac, V. (2014). *Pojava društvenih mreža kao globalnog komunikacijskog fenomena*. 206-219.
- [8] Horvatin, T. (2014). Retrieved from Ekonomski Portal, <https://ekonomskiportal.com/>
- [9] Ružić, D., Turkalj, D., & Biloš, A. (2014). *E-marketing*, III. izmijenjeno i prošireno izdanje. Osijek: Sveučilište Josipa Jurja Strossmayera u Osijeku, Ekonomski fakultet u Osijeku.
- [10] Supartiwi, H. A., Ariyanti, M., & Trianasari, N. (2023). The Effect of Social Media Marketing with Digital Platforms as the Communication Media of Brand Awareness on SME Products. In *Sustainable Future: Trends, Strategies and Development*.
- [11] Whiting, A., & Williams, D. (2013, August 30). *Why people use social media: a uses and gratifications approach*. Retrieved from Emerald Insight: <https://www.emerald.com/insight/content/doi/10.1108/QMR-06-2013-0041/full/html>

ARTIFICIAL INTELLIGENCE, PRIVACY AND INFORMATION SECURITY IN CRITICAL INFRASTRUCTURE SYSTEMS

Rade Dragović¹, Nadežda Ljubojev², Dalibor Dobrilović², Bojan Perović¹, Vladimir Milošev², Milada Novaković³

¹ Institute for standards and technology, Belgrade, Serbia

² University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

³ Technical College of Applied Sciences, Zrenjanin, Serbia

rade.dragovic@instate.biz

Abstract:

The accelerated technological development in the field of information technology has not been adequately accompanied by a legal procedural and organizational framework. Only the initial legal framework has been set for the area of information security, privacy and critical infrastructure, which in practice is not covered by organizational measures. An even greater discrepancy is present in attempts to apply artificial intelligence in critical infrastructure systems where there are no established good technological practices and the legislative framework only partially recognizes and regulates this area. The paper aims to present experiential recommendations in critical infrastructure systems for the areas of privacy and personal data protection, information security and artificial intelligence. Through the work, recommendations will be given for the establishment of an ISO-based management system that should meet the basic functional requirements related to critical infrastructure systems. The authors especially emphasize the need to apply information security and privacy in critical infrastructure systems, but at the same time they challenge the justification of applying artificial intelligence in critical infrastructure.

Keywords:

information security, artificial intelligence privacy, critical infrastructure

1. Introduction

The global digitization of all aspects of life is a thing of the past. The functioning of the state, economy and individuals is impossible without electronic services, which have become indispensable even in non-technologically based areas. Access to a large amount of data has been established, and at the same time, a legal or physical entity provides a large amount of data about itself. A large amount of data is processed, and data is collected without a clearly defined basis. From the collection of a large amount of data, the foundations for some advanced techniques have been created, the most popular of which is artificial intelligence. We are witnessing that many things are called artificial intelligence for the purposes of popularization, although in essence it is not. The most widespread technique for creating predictive statistical models from large amounts of data is machine learning. The popularization of artificial intelligence is followed by the media, politically, with disputed methodological bases of analysis, but also with great hope that it will improve all aspects of life and bring benefits to individuals, the economy and state bodies. In all these mostly disputed assumptions, it is necessary to pay attention to the technological, legal and organizational framework of the critical infrastructure, which must carefully apply the requirements of privacy and information security to consider the possibilities of applying artificial intelligence.

2. Critical infrastructure

Critical infrastructure is any system whose interruption of functioning or interruption of delivery of goods or services can have serious consequences on national security, health and lives of people,

property, environment, citizens' safety, economic stability, i.e. threatens the functioning of the state. Protection of critical infrastructure is a set of activities and measures aimed at establishing the functioning of critical infrastructure in the event of disruption or destruction, that is, protection in the event of threats and prevention of the consequences of disruption or destruction [1]. The sectors that are most often identified as critical infrastructure are: energy, transport, water and food supply, health, finance, telecommunications and information technologies, environmental protection and the functioning of state bodies.

The initial procedural framework for critical infrastructure is the relevant law, most often the Law on Critical Infrastructure and the standard ISO 22301 Security and resilience - Business continuity management systems - Requirements (international but also national standard) [2].

State bodies are usually charged by law to be guided by the following principles in the area of critical infrastructure protection: 1) the principle of an integrated approach - in the protection of critical infrastructure before, during and after disruptions or interruptions in the functioning of critical infrastructure, competent state authorities and reference entities participate, taking into account the various types of dangers arising from the risk analysis, and taking into account the interdependence of the critical infrastructure sector infrastructures and their interaction; 2) the principle of responsibility - the operators of the critical infrastructure are directly responsible for the functioning of the critical infrastructure, and for the improvement of the protection the critical infrastructure, in addition to the operators, all competent state bodies and reference entities are also responsible; 3) the principle of protection against various types of threats - operators, competent state authorities and reference entities in ensuring the continuous operation of critical infrastructure are obliged to take into account various types of risks; 4) the principle of continuous planning for the protection of critical infrastructure - the protection of critical infrastructure is based on a constant process of risk analysis for the functioning of critical infrastructure and assessment of the adequacy of protection measures; 5) the principle of data and information exchange and data protection - state authorities and reference entities are obliged to timely and continuously exchange the necessary data and information while protecting data related to critical infrastructure, in accordance with the regulations, regulating the protection of secret data.

In practice, acting according to the stated principles does not work in the prescribed way: 1) the principle of an integrated approach - competent state authorities and reference entities did not perform a risk analysis taking into account the interdependence of critical infrastructure sectors and their interaction or did not perform a risk analysis in the domain of critical infrastructure requirements at all; 2) the principle of responsibility - competent state authorities and reference entities are not obliged to document what their responsibilities are in the domain of critical infrastructure requirements; 3) the principle of protection against various types of threats - competent state authorities and reference entities did not carry out identification, analysis and guidelines for reacting to various types of risks; 4) the principle of continuous planning for the protection of critical infrastructure - competent state authorities and reference entities do not perform a continuous process of risk analysis for the functioning of critical infrastructure and do not perform assessments of the adequacy of protection measures; 5) the principle of data and information exchange and data protection - competent state authorities and reference entities do not perform timely and continuous exchange of necessary data and information while simultaneously protecting data related to critical infrastructure. They are not aware of the necessary data they need to handle nor what is the adequate time frame for certain data. The above examples show that the legal procedural framework cannot function by itself without integration with the missing technological and organizational framework.

The recommendation is to implement a legal procedural, technological and organizational framework that will establish multidisciplinary and multidepartmental requirements of the stated principles in practice by executing all requirements with the introduction of a body for coordination and a body for monitoring prescribed requirements. The best implementation path is the implementation of the required ISO standards specific to each critical infrastructure entity. When forming teams that will deal with the fulfillment of these principles, it is necessary to insist on a multidisciplinary approach with the external inclusion of missing expertise. The above guidelines are provided for the purpose of improving and fully implementing the missing functional framework so that the critical infrastructure can perform its designed and required functions.

3. Data privacy

The protection of personal data includes the rights of natural persons in relation to the processing of their personal data, the security of that data, as well as the obligations of the data handlers which own or process that data. Data privacy is predominantly legally procedural dimension, but for the purposes of full functionality. It must be integrated with the technological requirements of the implemented environment and organizational measures of human capacities in the organization [3].

The initial procedural framework for data privacy is the appropriate law, most often the Personal Data Protection Act (in European countries adopted on the model of the General Data Protection Regulation (GDPR) from 2016) and the standard ISO/IEC 27701 Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management - Requirements and guidelines (international but also national standard).

We are witnessing a society characterized by the processing of a large amount of different data in the *BIG DATA era*. Data on users, their activities and behavior are used for analysis and creation of user profiles, targeted marketing of products or services adapted to the personal characteristics and needs of users. The practice of data collection is sometimes clear, although not approved, for private companies that view data as a development resource, but it is unclear for state authorities that use monopolistically established services to collect and forward data to unknown sources for data processing. It is an open secret that Internet sites, even state ones, can function without Cookies (small files of information that a web server generates and sends to a web browser). Personal data is all data related to a person, based on which he can be identified (directly or indirectly) and his privacy can be threatened. Personal data means: Name and surname of a person, Residential address, e-mail address, Photo, IP address, Location where the person is located, Data used for analysis and creation of the user's profile (working abilities, economic situation, personal interests, behavior, consumer habits, movement, etc.) and the person's online behavior (data collected through web cookies).

Figure 1: An example 5W extended PIMS matrix

The EU General Data Protection Regulation - GDPR emphasized the protection of special categories of personal data: Racial affiliation, National affiliation, Religious affiliation, Trade union membership, Sexual orientation, Health information, Biometric data and Genetic data.

Full implementation of standards ISO/IEC 27701 Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management - Requirements and guidelines through the Privacy Information Management System (PIMS) is recognized by the establishment of all required

requirements in the standard and especially by the implementation of the 5W (WHO, WHY, WHERE, WHEN, WHAT) matrix with extended dimensions.

The author's recommendation is to implement the extended 5W matrix with the dimensions of access, risks and threats and vulnerabilities. The author's recommendation is to implement the extended 5W matrix with the addition of the dimensions of access, risks and threats and vulnerabilities. Identified data according to all the mentioned criteria represent the basis for establishing a privacy management system in a legal procedural, technological and organizational framework according to the requirements of the ISO/IEC 27701 standard.

4. Information security

Information security is a set of measures that enable the data handled by the application of the ICT system to be protected from unauthorized access, as well as to protect the integrity, availability, authenticity and non-repudiation of that data, so that the system functions as intended, when intended and under the control of authorized persons. The initial procedural framework for information security is the appropriate law, most often the Law on Information Security and the standard ISO/IEC 27001 Information security, cybersecurity and privacy protection - Information security management systems - Requirements. Information Security Management System (ISMS) should represent a systematic approach for establishing, implementing, functioning, monitoring, reviewing, maintaining and improving the organization's information security, in order to achieve business goals, but risk management must not be neglected [4]. The ISMS must be based on the assessment of information security risks and the level of acceptability of such risks by the state authority in such a way as to effectively and efficiently treat the risks and to manage the risks in an adequate manner [5]. Requirements for the protection of information assets, whether they are legal, regulatory, contractual or as a consequence of risk management and the application of appropriate controls, when necessary, contribute to the successful implementation of ISMS. The framework for the treatment of information security risks is defined through ISO/IEC 27005 Information security, cybersecurity and privacy protection - Guidance on managing information security risks.

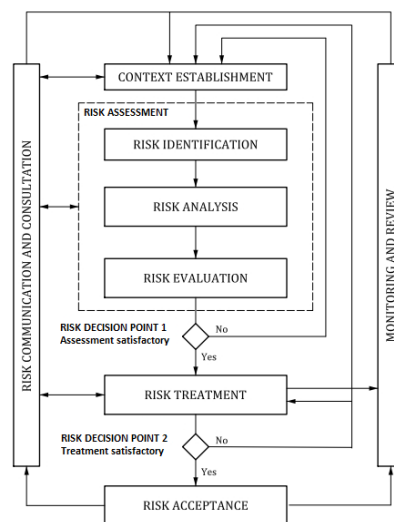


Figure 2. Information security risk management proces, ISO/IEC 27005

In addition to the 245 Published standards of the ISO Technical Committee ISO/IEC JTC 1/SC 27 - Information security, cybersecurity and privacy protection of which only 80 requirements the ISO/IEC 27K mark, the standard stands out in the topic of the work: ISO/IEC TR 27563 Security and privacy in artificial intelligence use cases - best practices. Standard outlines best practices on assessing security and privacy in artificial intelligence use cases, covering in particular those published in ISO/IEC TR 24030. The following aspects are addressed: an overall assessment of security and privacy on the AI system of interest; security and privacy concerns; security and privacy risks; security and privacy

controls; security and privacy assurance; and security and privacy plans. Security and privacy are treated separately as the analysis of security and the analysis of privacy can differ.

5. Artificial intelligence

Today, when many new disciplines by using artificial intelligence have developed, perhaps the definition of the ICT research house Gartner is the closest to today's moment: "Artificial intelligence (AI) applies advanced analysis and logic-based techniques, including machine learning, to interpret events, support and automating decisions and taking actions".

Artificial intelligence systems rely on data inputs from initial development through the training phase and use in real time. Considering the broad definition of personal data under European data protection laws, the development and use of AI systems will often result in the processing of personal data. Artificial intelligence is a regulated technology through EU AI law, so practitioners and organizations must understand the close relationship between personal data protection law, information security and EU AI Act. The European Union's approach to artificial intelligence is based on the idea that individuals and companies use the advantages of artificial intelligence, but in a safe way and with the existence of precise rules, in order to ensure the smooth functioning of the market and the public sector and to ensure the protection of the fundamental rights of individuals [6]. The General Data Protection Regulation ("GDPR") is a technology-neutral regulation. As the definition of "processing" under the GDPR is broad (and in practice includes nearly all activities conducted on personal data, including data storage), it is evident that the GDPR applies to AI systems, to the extent that personal data is present somewhere in the lifecycle of an AI system. It is very difficult to separate personal data from non-personal data, which increases the likelihood that AI systems process personal data at some point within their lifecycle. While AI is not explicitly mentioned in the GDPR, the automated decision-making framework (article 22 GDPR) serves as a form of indirect control over the use of AI systems, on the basis that AI systems are frequently used to take automated decisions that impact individuals. The AI Act has established different rules for different levels of risk. The new rules establish obligations for providers and users depending on the level of risk from artificial intelligence. While many AI systems pose minimal risk, they need to be assessed. AI systems that negatively affect safety or fundamental rights will be considered high risk and will be divided into two categories:

- I) AI systems that are used in products falling under the EU's product safety legislation.
- II) AI systems falling into specific areas that will have to be registered in an EU database:

- **Management and operation of critical infrastructure**

- Education and vocational training
- Employment, worker management and access to self-employment
- Access to and enjoyment of essential private services and public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Assistance in legal interpretation and application of the law.

The reference standard for artificial intelligence is ISO/IEC 42001 Information technology - Artificial intelligence - Management system. In addition to the above standard, the technical committee ISO/IEC JTC 1/SC 42 - Artificial intelligence published 31 standards, of which special attention should be paid to the following standards:

- ISO/IEC 5339 Information technology - Artificial intelligence - Guidance for AI applications.
- ISO/IEC TR 5469 Artificial intelligence - Functional safety and AI systems.
- ISO/IEC TS 8200 Information technology - Artificial intelligence - Controllability of automated artificial intelligence systems.
- ISO/IEC 23894 Information technology - Artificial intelligence - Guidance on risk management.
- ISO/IEC TR 24027 Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making.
- ISO/IEC TR 24028 Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence.
- ISO/IEC TR 24368 Information technology - Artificial intelligence - Overview of ethical and societal concerns.

- ISO/IEC 38507 Information technology - Governance of IT - Governance implications of the use of artificial intelligence by organizations.

Special attention of the future vision of creating requirements of future standards is caused by the withdrawal of standards ISO/IEC WD TS 22440 Artificial intelligence - Functional safety and AI systems - Requirements in phase 20.98 (20 Preparatory, after the adopted phase 60 Completion of main action). ISO/IEC WD TS 22440 provides requirements and guidance on the terminology, properties, risk factors, processes, methods, techniques and architectures relating to: - use of AI technology within a safety-related function; - use of safety-related function based on conventional technology to ensure safety of a system using AI technology; - use of AI technology to design, develop and verify safety-related functions. This document includes general considerations on how security threats can affect safety of an AI system. Unless differently specified, this document is applicable to all types of AI technologies. It includes specific details on machine learning (ML). A very important topic of standards: Artificial intelligence - Functional safety and AI systems - Requirements strategically continues the development of standards ISO/IEC AWI TS 22440-1 Artificial intelligence - Functional safety and AI systems - Part 1: Requirements, ISO/IEC AWI TS 22440-2 Artificial intelligence - Functional safety and AI systems - Part 2: Guidance and ISO/IEC AWI TS 22440-3 Artificial intelligence - Functional safety and AI systems - Part 3: Examples of application.

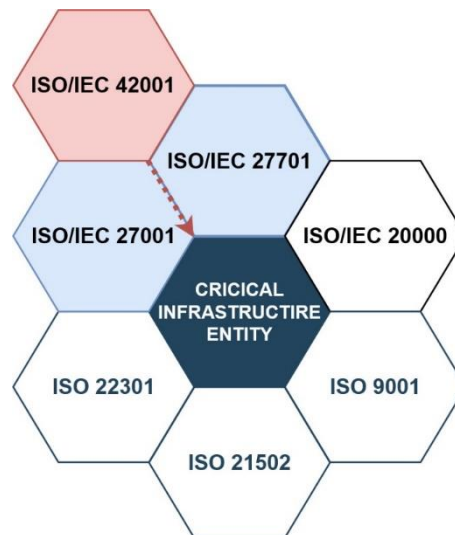


Figure 3. Relationship between basic ISO standards of critical infrastructure and artificial intelligence

The basic standards of critical infrastructure are ISO/IEC 27001 Information security, cybersecurity and privacy protection - Information security management systems - Requirements, ISO/IEC 27701 Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management - Requirements and guidelines, ISO/IEC 20000-1 Information technology - Service management Part 1: Service management system requirements, ISO 22301 Security and resilience - Business continuity management systems - Requirements, ISO 9001 Quality management systems - Requirements, ISO 21502 Project, programme and portfolio management - Guidance on project management. The requirements of the mentioned standards are guidelines for the possible inclusion of the Artificial Intelligence System according to the standard ISO/IEC 42001 Information technology - Artificial intelligence - Management system.

It is possible to use artificial intelligence, only if it is procedurally arranged according to the requirements of standards (international and mostly national) ISO/IEC 42001 Information technology - Artificial intelligence - Management system. Application of the specified standard in the system of critical infrastructure can be established only if the requirements of ISO/IEC 27001 Information security, cybersecurity and privacy protection - Information security management systems - Requirements and ISO/IEC 27701 Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management - Requirements and guidelines.

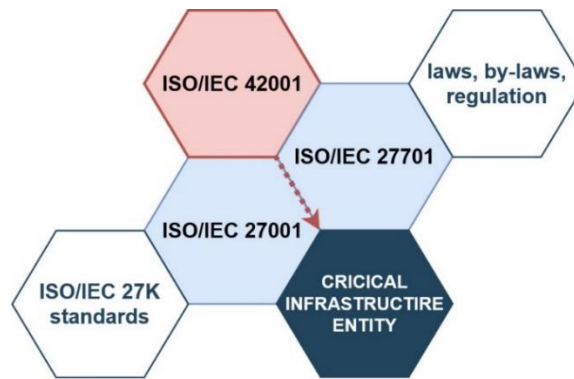


Figure 4. Relationship between ISMS and PIMS of critical infrastructure and AI

The requirements of the ISO/IEC 27701 standard are extended to the requirements of the standard in the field of personal data protection, European and national regulations in the field of privacy, while the requirements of the ISO/IEC 27001 standard are extended to the relevant standards from the family of the ISO/IEC 27K standard, but also to specific functional links entities of critical infrastructure in the areas of standards that they apply in their work.

The author's recommendation is that of the 80 published standards of the ISO technical committee ISO/IEC JTC 1/SC 27 - Information security, cybersecurity and privacy protection which bear the ISO/IEC 27K mark, attention is required in addition to the base standard ISO/IEC 27001 Information security, cybersecurity and privacy protection - Information security management systems - Requirements, conform to the following standards:

- ISO/IEC 27018 Information technology - Security techniques - Code of practice for protection of personally identifiable information (PII) in public clouds acting as PII processors,
- ISO/IEC 27021 Information technology - Security techniques - Competence requirements for information security management systems professionals,
- ISO/IEC 27036-2 Cybersecurity - Supplier relationships - Part 2: Requirements,
- ISO/IEC 27036-3 Cybersecurity - Supplier relationships - Part 3: Guidelines for hardware, software, and services supply chain security,
- ISO/IEC 27037 Information technology - Security techniques - Guidelines for identification, collection, acquisition and preservation of digital evidence,
- ISO/IEC 27039 Information technology - Security techniques - Selection, deployment and operations of intrusion detection and prevention systems (IDPS),
- ISO/IEC 27042 Information technology - Security techniques - Guidelines for the analysis and interpretation of digital evidence,
- ISO/IEC 27070 Information technology - Security techniques - Requirements for establishing virtualized roots of trust.

Analyzing the stated requirements of the standard, we come to the conclusion that the implementation of artificial intelligence in critical infrastructure systems is not only a matter of trend and the desire to facilitate ongoing work, but is a risk-based business that opens up many dilemmas [7]. The work of the critical infrastructure entity must not be viewed only from the technological aspect while neglecting the legal procedural and organizational part of the requirements. In practice, it has been shown that systems that emphasize a low price cannot achieve a sufficient level of acceptable risk, and that they are inversely proportional. The level of information security increases with the price of implemented solutions, with the emphasis that a high price does not guarantee a high level of information security. The critical infrastructure entity must define an acceptable level of risk that should result from the treatment of individual risks that are specific to that entity as required by the standards. There is no technology that can violate the conclusions of risk treatment and be positioned outside the zone of acceptable level of risk. Globally based AI systems from the aspect of risk analysis show that the application of such concept is unacceptable from the aspect of information security. State-based and Entity-based AI systems are acceptable from the point of view of information security risk analysis, but they represent a huge investment that most countries in Europe are unable to finance. There are proposals to build the concept of State based AI system as EU member-based AI system for critical infrastructure systems of EU member states.

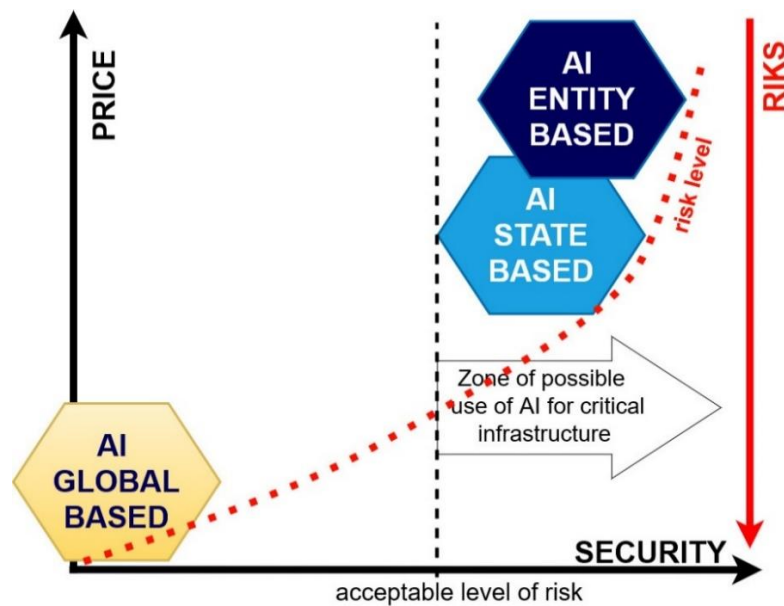


Figure 5. Zone of possible use of AI for critical infrastructure systems

6. Conclusions

The use of artificial intelligence systems is a modern trend that should not impress critical infrastructure systems in the hope that it will solve all their problems, but the eventual use of artificial intelligence should be based exclusively on risk management. The guidelines for the above conclusion were given on the basis of numerous analyzes in the critical infrastructure system. The above conclusion is a requirement of the legal procedural framework of information security and privacy protection, and certainly represents a technological gap in critical infrastructure systems. In critical infrastructure systems, it must be insisted that decisions about the use of artificial intelligence are made only on the basis of risk treatment.

References:

- [1] Dragović, R. Ivković, J., Dragović, D., Klipa, Dj., Decision support system as a support for strategic management of state administration (YUINFO 2015) Proceedings, ISBN: 978-86-85525-15-5
- [2] Dragović, R., Dobrilovic, D., & Dragović, D. (2022). Recommendations for the Creation of Usable Critical Infrastructure for the Delivery of Priority Services of State Bodies. In IFIP International Conference on Human Choice and Computers (pp. 161-171). Cham: Springer Nature Switzerland. DOI:10.1007/978-3-031-47990-8_15
- [3] Dragovic, R., Ivkovic, M., Perovic, B., & Klipa, D. Dataveillance and data mining as a technology support to the process of investigation (TELFOR 2011) Proceedings of Papers (pp. 1257-1260). IEEE., 2011, DOI: 10.1109/TELFOR.2011.6143780
- [4] Dobrilovic D. and Dragović R. Usage of Threat Modelling Tools for Implementation of ICT Security Standards in IoT Environments (ICCS-DE 24) ISBN: 978-5-6041814-6-1
- [5] Dragović, R., Stanisavljev, S., Dobrilovic, D., Dragović, D., Milosev, V., Software information security management for government authorities, IIZS 2023, ISBN 978-86-7672-368-3
- [6] Ljubojev, N. P., & Dukic-Mijatovic, M. (2019). The Right of Interactive Communication of the Work to the Public. Kultura Polisa, 16, 683.
- [7] Dragović, R., Makitan, V., Dobrilović, D., Perović, B., & Kostovski, I. (2024). Operational research in function of maintaining an information security risks of critical infrastructure, Engineering management and competitiveness, EMC 2024, ISBN: 978-86-7672-374-4

Programming model for GenICam-based cameras

Bojan Torbica¹, Dragan Ivetić¹ and Aleksandar Kupusinać¹

¹ Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia

bojan.torbica@outlook.com; ivetic@uns.ac.rs; sasak@uns.ac.rs

Abstract:

The proliferation of machine vision systems has driven the need for efficient management and control of single, or multiple cameras, particularly in industrial and research settings where reliability and precision are paramount. This paper presents a programming model tailored for GenICam-based cameras and aimed at abstracting the complexities associated with managing a large number of cameras within a unified, programming interface. Key challenges addressed include ensuring reliable image capture and providing a scalable solution adaptable to varying hardware setups. The implementation of this model can help users achieve consistent image acquisition performance, minimize programming overhead, and facilitate easier integration into larger vision systems. Carefully assessing the need of the domain in terms of image data capture is essential in determining the type of data structures needed and the way the image data itself is meant to be used.

Keywords:

genicam, machine vision cameras, machine vision, programming model

1. Introduction

Over the last several decades, digital camera device technology has made huge leaps in quality and accessibility, both to consumers and professional users [1]. The area where this is most visible is in the mobile phones industry, as the rise of social media presence has increased the demand for higher quality imaging devices [2]. These devices continue to improve the number of pixels available, as well as the general quality of images they generate. Mobile phone cameras generally provide their users with control that is of a much coarser grain than some professional environments might need. For these, depending on the concrete application, different types of camera devices are suited. For professional photography of different styles, SLR (Single-Lens Reflex) type camera and ILC (Interchangeable Lens Cameras) are dominant.

In a large majority of industrial applications, machine vision cameras are used for their speed, level of control and reliability in operation. These industrial applications include, automotive industry, food and beverage industry, pharmaceutical industry, scientific research, etc. [3,4] In these industries, the importance of consistency, reliability and correctness is imperative. This is combined with the need for a software suite that will reliably deliver image information and process it accordingly. For applications requiring consistency in operation and precision, it is often both convenient and cost effective to have generic software control for many types of devices from different vendors. The GenICam (Generic Interface for Cameras) standard was created to, in a way, solve this particular problem [5].

Machine vision cameras come in a wide array of types [6], each tailored to specific tasks within industrial and scientific applications. Area scan cameras are among the most common, used for capturing 2D images in applications such as quality inspection and object recognition. Line scan cameras, on the other hand, are ideal for continuous processes like web inspection and high-speed sorting, where a single line of pixels is captured repeatedly to form a complete image. Thermal imaging cameras detect infrared radiation and are used in applications requiring temperature measurement, such as monitoring industrial equipment or detecting heat leaks. Hyperspectral cameras capture images across a wide spectrum of light beyond visible wavelengths, providing detailed information for applications in agriculture, medical diagnostics, and mineral exploration. Additionally, each type of

camera model mentioned here, comes in a wide variety of hardware interfaces. These include Ethernet-based cameras, USB cameras, FireWire cameras, CoaXPress, etc., each with its own challenges in terms of system integration, requiring different hardware support. All this deepens the potential complexity of vision systems and makes it difficult for system integrators to make solid, future-proof decisions regarding their vision application, further emphasizing the need for a solid, all-encompassing software solution, for all machine vision camera types.

Though not a novel idea [7] in and of itself, the goal of this paper is to present a data model to be used for quickly building image capture applications for GenICam based devices by users that have some "hands-on", but not necessarily much programming experience in and of itself. The idea is to have fast iteration on application development, providing fast turnaround and immediate feedback on the application itself. This model is designed to be as simple as possible while being generic enough to cover a wide range of use cases, mostly focused on simple data capture. This programming model can be used as a base to develop more complex pieces of software depending on the need of the application in question. The accent here is in the generic nature of the data structures given in the model. In the paper, following this chapter, we will give a brief overview of the GenICam standard and how it is used. Then, we will present the camera model which we will refer to for programming. After that, we present the programming model itself and the accompanying API for it that can be used to build more complex software. Lastly, we will give a conclusion in which we reflect on the programming model presented and present some possible future work that can be done based on it.

2. GenICam

Today's digital cameras, in general have much more functionality than just simply delivering an image. Processing the image and appending the results to the image data stream, controlling external hardware, etc. have become common tasks for machine vision cameras. As a result, the programming interface for cameras has become more complex. The goal of GenICam is to provide a generic programming interface for all kinds of cameras. No matter what interface technology the cameras are using or what features they are implementing, the application programming interface (API) should be always the same.

The GenICam standard consists of multiple modules [8]:

- GenApi: Defines the format of an extensible mark-up language (XML) file called camera description file that contains all of the required information to automatically map a camera's features to its registers. Adding a new feature to a camera means simply extending the camera's description file, thus making the new feature immediately available to all GenICam aware applications.
- Standard Features Naming Convention (SFNC): Standardizes the names of more than 220 commonly used camera features. To comply with GigE Vision and USB3 Vision, only a handful of the features such as PixelFormat, etc. are considered mandatory. The rest are either recommended or optional. Compliance with the naming convention is important for interoperability, as it frees application software from the complexity of situations where vendors call the same feature by different names, such as, 'Brightness' and 'Gain'.
- Defines a software interface provided by some manufacturers of traditional frame grabbers. The purpose of GenTL is to provide a generic way to enumerate devices, communicate with devices and, if possible, stream data from devices to the host in a manner independent from the underlying transport technology. This allows third-party software (such as our proposed programming model) to use different technologies to control cameras and to acquire data in a transport layer agnostic way.
- GenDC: Defines a Generic Data Container (GenDC) format, which allows devices to send any form of data (including 3D, multispectral, metadata etc.) in the TL protocol, independent of its format.
- GenCP - Hardware independent protocol which defines register access and events.
- FWUpdate - Specifies a format for a Generic Update File (GUF). Such an update file can contain multiple firmware updates for an arbitrary number of devices.

- CLProtocol – Allows cameras that comply with the Camera Link standard to be accessed through GenApi.

Since the standard by design is meant to be modular, each of the previously mentioned modules can be released independently from each other.

Detailing each of the modules is beyond the scope of this paper, however the reader is highly encouraged to read through the EMVA's documentation regarding the standard [8].

3. Camera model framework

In this section we introduce a model for a camera device that is based and implemented using the GenICam standard.

We will define our camera as an imaging device consisting of a single-color imager (sensor), a central control unit in the form of a microcontroller that is in charge of controlling camera features and sensor read out and transfer to the Host. Host in this case refers to a personal computer to which the camera device is connected to. Additionally, we will define camera features as user-controllable settings on a camera. The exact architecture and implementation of said features (and consequently the microcontroller) is not relevant and beyond the scope of this paper; the hardware is presented to give a sort of frame of reference for our API. Following the GenAPI model, features will be divided into several categories. Categories can be understood as collections of features that have a similar role in what they control in a camera. Categories we will define are as follows: Device, Format, Acquisition, Analog, and User. Each of the categories has several features within it. Features are divided further by their type. Types of features are Integer, Float, Boolean, String and Enumeration.

The GenICam standard itself defines very few mandatory features that all device implementations must have. However, for the purpose of having a functional camera device, some features must be declared. In this paper, we will consider a device that has the following features defined:

- WidthMax - Maximum width of the image (in pixels). The dimension is calculated after horizontal binning, decimation or any other function changing the horizontal dimension of the image.
- HeightMax - Maximum height of the image (in pixels). This dimension is calculated after vertical binning, decimation or any other function changing the vertical dimension of the image.
- Width - Describes the width of the resulting image.
- Height - Describes the height of the resulting image.
- PixelFormat - Describes the format of pixels provided by the device. Describes both the size of the pixels and color filter, if applicable [9].
- ExposureTime - Describes how long (in units of time, defined by the implementation) the sensor will be exposed to light during one cycle of image acquisition.
- TriggerSource - Describes how the image acquisition is started. This is usually either by using a software generated signal, or a hardware (logic) signal.
- TriggerSelector - Describes how many triggers are expected in acquisition: FrameStart, each trigger signal will start capture of a single frame, and AcquisitionStart, where only one trigger signal is expected to start frame capture (this is usually called a free-run mode).
- AcquisitionStart - Command to start the acquisition engine in the device.
- AcquisitionStop - Command to stop the device's acquisition in the device.

The features, as defined here, are based on the GenICam SFNC [10].

4. Camera control model

This section will describe a control model for controlling GenICam-based cameras. In order to control GenICam-based cameras we must first load the Provider given by a Vendor. Providers are runtime binaries that are loaded by the client application. They provide implementation for function pointers defined in the GenTL API which the client calls. For the purposes of this paper, the code for loading the runtime library will not be shown as it is beyond the scope of this paper. It can be assumed that appropriate code is written to load the Provider and that our API functions make use of Provider functions appropriately. The programming model and API we present in this paper will be written in the C language. The reason for choosing C is due to its versatility, stable ABI and the ability to quickly and efficiently write bindings to any other language.

The entry point of our programming API must be a function to discover cameras currently connected to the Host. This function is defined as follows:

```
cam_err cam_discovery(int* camera_count);
```

The discovery function returns through the output argument the number of devices currently connected to the Host. As each function in the API returns an error code from which the user can determine the cause of the error. In order to facilitate effective error reporting, two additional functions are added:

```
void cam_last_err_code(cam_err* err_code);
const char* cam_last_err_msg();
```

After discovery returns a number of cameras greater than 0, the user is expected to “open” the camera. The open procedure is defined as a sequence of calls to the device that will initialize it and prepare it for user control. This procedure depends on the device itself and it is abstracted away from the client of the GenICam device.

```
cam_err cam_open(int i, cam_camera* cam);
```

The function accepts an index into the internal list of devices and as output it populates the user’s structure representing the camera device. The structure returned as an output argument to the function contains all the information about the device, including its model name, serial number, features, etc.

Camera features can have values of different types. For this purpose, the definitions of a camera feature can use a C union to represent these values. Based on the type, clients can then read the appropriate value. Outside of their name, type and value, features are defined by their access mode. Access mode describes the feature’s usability in terms of whether it can only be read from, written to, or both read from and written to. A feature can be updated, or queried by using the following functions:

```
cam_error cam_get_feature(cam_camera* c, const char* n, cam_feature* f);
cam_error cam_set_feature(cam_camera* cam, cam_feature* ftr);
```

After opening the camera, the user must have the option to further control the device by starting and stopping acquisition and closing the camera once it is no longer needed. Our API provides appropriate functions:

```
cam_err cam_start_acq(cam_camera* cam, cam_callback func);
cam_err cam_stop_acq(cam_camera* cam);
cam_err cam_close(cam_camera* cam);
```

We will define the callback function that is passed as an argument to acquisition start in the following way:

```
typedef void (*cam_callback) (cam_err error, cam_img* image, void* usr_d);
```

Finally, the structure to hold the image data and metadata (additional data related to the image payload) should contain information such as width and height of the image, number of channels the image contains (for example, the data can be a color image and therefore have 3 channels), the pixel format that describes the data bytes, etc.

Previously described structures and defined functions are minimally required to be able to efficiently control a GenICam-based camera. This API can also be applied to any bespoke SDK for camera application development, by any vendor. However, this is not always efficient as each SDK is purpose built for and by a specific vendor. Thus, this potentially locks the third-party application to that specific vendor. Making a switch to a different vendor could prove costly in such a situation. Having one exposed API, like we have shown here, and basing its implementation on the GenICam standard, ensures a vendor-agnostic way of camera control. Additionally, it can be assumed that the performance of acquisition is as best as possible; this means that if we are set up to capture 60 frames per second, we will capture at that frame rate, as stable as possible on a given platform. In highly sensitive applications, critical code paths must be kept as short as possible. In such applications it is usual to have multi-threaded code and data protection becomes essential. For example, it is common (and useful) to have a separate thread from the main one open for image data acquisition and processing. This thread might hold data necessary for the rest of the application thus necessitating a form of synchronization and data protection. The reader can assume that the API itself is implemented in a thread-safe manner.

5. Use case

In this section we describe a use case that requires an application which makes use of our API. The application must discover and open cameras, start acquisition, acquire images in accordance with the use case, stop acquisition and close the camera to clean up any internally allocated resources. For the sake of simplicity, we can assume that all calls are successful. Production code is highly advised to have proper error checking.

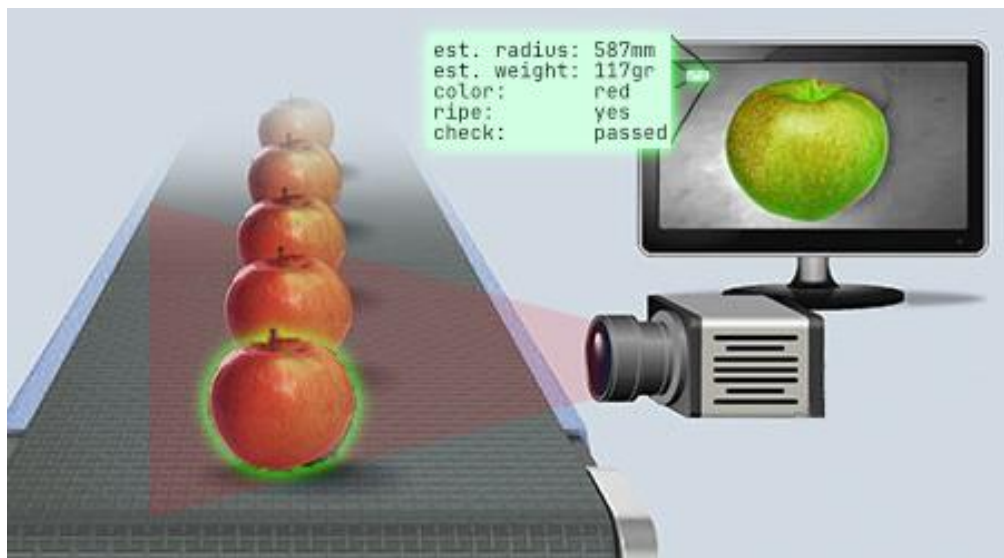


Figure 1: Example of the use case setup we are considering. Inspection shows dimension and weight estimates as well as ripeness and if the current item passes inspection. Image edited from [11].

The use case we will be considering here is a packaging inspection line in a food production facility. Figure 1 is a sketch of the setup for the use case [11]. The setup is such that the food packaging is traveling through the conveyor belt, and a camera is placed orthogonally against the belt and aimed

directly at the passing packages. This camera is taking images of the packages. The control software, built with our programming model, performs a check based on some machine vision algorithm which determines whether or not the packaging is good. The camera is set to receive a hardware trigger and capture only one frame per trigger. The trigger is activated as a package that passes through a sensor which activates a light source with a hardware trigger for the camera.

As mentioned before, the application we create must begin with camera device discovery. Following this basic discovery, we will continue preparing for acquisition. For the purposes of this use case, we will assume that only one camera is in the system. We then set two features, exposure time and pixel format. These values should be set according to real world conditions and considerations. For example, exposure time should be kept at a reasonable minimum in order to avoid problems such as motion blur [12]. Motion blur can happen in environments where objects are moving, such as in our use case. Pixel format should be chosen based on the machine vision algorithm needed to complete the inspection. Ideally, formats that have raw pixel values should be preferred since they carry data that is not altered in any way allowing for machine vision applications and algorithms to have true representation of the captured environment. After everything is set up, we start the acquisition and let the callback function perform all the necessary functionality needed for our use case. After the system is no longer needed, we “shutdown” the camera and cleanup resources.

6. Conclusions

In this paper a sample of data structures and code examples is given to illustrate and virtualize a real-world example of an application created to capture images from a camera. The camera model given here is intentionally left as simple as possible in order to cover the basics of what a camera might represent, programmatically speaking. Though we implement the API in a thread-safe manner, the clients of the API must be careful and design their applications with thread-safety in mind.

Vendors will often use the "looseness" of the standard to implement certain features in a way not compatible with other vendors, or even camera models from the same vendor. Thus, clients (in this case, us with our API) have to adjust the implementation in order to fit the specificity of the vendor implementation of the standard. As an example, one vendor might decide to forbid clients from giving allocated buffers for the data payload to the acquisition engine. Therefore, we would have to adjust the implementation in our API in order to distinguish between cases where we can or cannot commit allocated data buffers. However, even in this, the GenAPI/GenTL provides a way to query such a capability. Additionally, vendors do not usually document their GenICam implementations as well as their bespoke SDKs. All this, in a way, implies that in order to truly create a robust API and application for control of GenICam-based devices we unfortunately require some trial and error. Aforementioned is the personal and professional experience of the author.

Acknowledgment:

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad” (No. 01-3394/1).

References:

- [1] Y. Lu, S. Liu, and Y. Bai, “Analysis of digital photography technology in the era of big data,” *Mobile Information Systems*, vol. 2022, pp. 1–8, 06 2022.
- [2] C. Peters and S. Allan, “Everyday imagery: Users reflections on smart-phone cameras and communication,” *Convergence: The International Journal of Research into New Media Technologies*, vol. 24, 11 2016.

- [3] S. S, “Industrial application of machine vision,” International Journal of Research in Engineering and Technology, vol. Volume: 03, 05 2014.
- [4] E. Malamas, E. Petrakis, M. Zervakis, L. Petit, and J.-D. Legat, “A survey on industrial vision systems, applications and tools,” Image and Vision Computing, vol. 21, pp. 171–188, 02 2003.
- [5] M. Schmidt. (2023, 11) GenICam: In the Engine Room of Machine Vision. URL: <https://www.vision-systems.com/cameras-accessories/article/14301062/genicam-in-the-engine-room-of-machine-vision>
- [6] M. Javaid, A. Haleem, R. P. Singh, S. Rab, and R. Suman, “Exploring impact and features of machine vision for progressive industry 4.0 culture”, Sensors International, vol. 3, p. 100132, 2022. URL: <https://www.sciencedirect.com/science/article/pii/S266635112100053X>
- [7] A3. (2024, 07) egrabber: A single api for gige vision, coaxpress camera link cameras. URL: <https://www.automate.org/vision/tech-papers/egrabber-a-single-api-for-gige-vision-coaxpress-and-camera-link-cameras>
- [8] EMVA. (2024) Introduction. URL: <https://www.emva.org/standards-technology/genicam/introduction-new/>
- [9] EMVA. (2021, 06) Pixel format naming convention. URL: https://www.emva.org/wp-content/uploads/GenICam_PFNC_2_4.pdf
- [10] EMVA. (2022, 02) Standard features naming convention. URL: https://web.archive.org/web/20240812193902/https://www.emva.org/wp-content/uploads/GenICam_SFNC_v2_7.pdf
- [11] A. Tu. (2022, 06) Assessing food inspection techniques. URL: https://www.photonics.com/Articles/Assessing_Food_Inspection_Techniques/a68120
- [12] H. Lin, D. Mullins, D. Molloy, E. Ward, F. Collins, P. Denny, M. Glavin, B. Deegan, E. Jones, “Optimizing Camera Exposure Time for Automotive Applications”, Sensors 2024.

Security of Information Systems in Solar Energy: Overview of Approaches and Practices

Luka Djordjević¹, Borivoj Novaković¹, Mihalj Bakator¹, Velibor Premčevski¹ and Stefan Ugrinov¹

¹ University of Novi Sad / Technical Faculty "Mihajlo Pupin", Đure Đakovića St., 23000 Zrenjanin, Serbia

luka.djordjevic@tfzr.rs;

borivoj.novakovic@tfzr.rs;

mihalj.bakator@tfzr.rs;

velibor.premcevski@tfzr.rs; stefan.ugrinov@tfzr.rs

Abstract:

As solar energy systems become an essential part of sustainable energy solutions, integrating them with digital technologies brings new security challenges. This paper explores the main vulnerabilities within solar information systems, such as cyber threats, physical security risks, and issues from outdated software. It focuses on implementing practical security practices, including encryption, multi-factor authentication, and access control, and examines how emerging technologies like IoT, blockchain, and AI can strengthen these defenses. The paper also underscores the role of regulatory frameworks and industry standards, which guide effective security practices and help ensure solar systems remain resilient and compliant in the face of threats. In conclusion, the findings emphasize that a layered security approach - one that combines proven methods, innovative technologies, and a readiness to adapt to new cyber challenges - is essential to safeguarding the reliability and security of solar energy system.

Keywords:

Solar energy security, cybersecurity, emerging technologies

1. Introduction

The global transition towards renewable energy sources has placed solar energy at the forefront of efforts to reduce carbon emissions and combat climate change [1]. Solar energy systems, once considered a niche technology, have evolved into a critical component of the modern energy landscape [2]. As these systems proliferate and become more integral to national and international energy grids, the reliance on advanced information systems to manage and optimize their performance has increased exponentially. This reliance, however, comes with a set of unique security challenges that must be addressed to ensure the reliability, efficiency, and safety of solar energy production and distribution.

Solar energy systems today are not merely collections of photovoltaic panels; they are complex, interconnected networks that include a variety of hardware and software components. These systems collect, process, and transmit large volumes of data to monitor performance, predict maintenance needs, and optimize energy output. This digital infrastructure is vital for the efficient operation of solar energy systems, particularly in the context of smart grids and IoT (Internet of Things) applications, where real-time data exchange and decision-making are crucial [3], [4].

However, the integration of these advanced information technologies also introduces significant vulnerabilities. Cybersecurity threats, ranging from data breaches to more sophisticated attacks such as ransomware or state-sponsored hacking, pose a real risk to the stability and integrity of solar energy systems. A successful cyber-attack on a solar energy system could lead to disruptions in energy supply, substantial financial losses, and even physical damage to critical infrastructure [5]. Given the essential role that solar energy plays in the broader energy ecosystem, ensuring the security of these systems is not only a technical necessity but also a matter of national and global security.

In addition to external cyber threats, internal challenges such as inadequate access controls, outdated software, and insufficient physical security measures can also undermine the security of solar energy systems [6]. As these systems often operate in distributed environments, securing the communication

channels between different components - whether they are on-site, in remote locations, or in the cloud - becomes increasingly complex.

This paper aims to provide a review of the security challenges inherent in the information systems that underpin solar energy infrastructures. It will explore the current landscape of vulnerabilities, analyze the effectiveness of existing security practices, and discuss emerging technologies that promise to enhance the protection of these critical systems. By offering a detailed examination of these issues, this review seeks to contribute to the development of more robust security strategies that can safeguard solar energy systems against both current and future threats. Through this exploration, the paper will also highlight the importance of regulatory frameworks and industry standards in guiding the implementation of best practices in solar energy security.

2. Overview of Security Challenges in Solar Energy Systems

As solar energy systems evolve and integrate more deeply into the digital landscape, the security challenges they face become increasingly complex and multifaceted. These challenges arise from the inherent vulnerabilities in their information infrastructure, as well as from the diverse array of cyber threats that target these systems. Understanding these challenges is crucial for developing effective security measures that can protect solar energy systems from potential disruptions and attacks.

2.1. Information Infrastructure Vulnerabilities

The information infrastructure of solar energy systems is composed of various interconnected components, including sensors, controllers, communication networks, and data storage systems [7]. Each of these components plays a critical role in ensuring the smooth operation and optimization of energy production. However, the complexity and distributed nature of this infrastructure introduce several vulnerabilities that can be exploited by malicious actors [8].

One of the primary vulnerabilities lies in the communication networks that connect different parts of the solar energy system. These networks, often relying on wireless communication protocols, are susceptible to interception, jamming, and spoofing attacks. For instance, attackers could intercept data transmissions between solar panels and the central control system, potentially altering the data to mislead operators or disrupt the system's performance. Additionally, unencrypted communication channels can expose sensitive data, making it easier for attackers to gain unauthorized access to the system.

Another significant vulnerability is found in the software that controls and monitors solar energy systems. Many of these systems rely on custom or proprietary software, which may not always be thoroughly tested for security flaws. Inadequate software security can lead to vulnerabilities such as buffer overflows, SQL injection attacks, and other forms of exploitation that can compromise the entire system. Moreover, outdated software or firmware, which often remains unpatched due to the decentralized nature of solar installations, presents a significant security risk, as known vulnerabilities can be easily exploited by attackers [9], [10].

Physical security is also a critical concern for the information infrastructure of solar energy systems. Many solar installations are located in remote or hard-to-reach areas, making them difficult to secure physically. Without adequate physical security measures, such as tamper-proof enclosures, surveillance, and access control systems, attackers could gain direct access to critical components, allowing them to sabotage equipment, steal data, or install malicious devices.

2.2. Cyber Threats and Risks

In addition to the vulnerabilities inherent in the infrastructure, solar energy systems face a wide range of cyber threats that can disrupt their operation and compromise their security. These threats are becoming increasingly sophisticated, driven by the growing interest in exploiting the weaknesses of critical infrastructure.

One of the most prominent cyber threats to solar energy systems is the risk of ransomware attacks. In such an attack, malicious software encrypts the data or system controls, rendering them unusable until a ransom is paid. Given the essential nature of solar energy in many regions, the impact of such an attack could be devastating, leading to significant downtime and financial losses. Moreover, ransomware attacks often target the most vulnerable systems, such as those with outdated or unpatched software, highlighting the importance of regular updates and maintenance [11], [12].

Another significant cyber threat is the potential for state-sponsored attacks, where adversaries target solar energy systems as part of broader geopolitical strategies. These attacks are typically highly sophisticated, involving advanced persistent threats (APTs) that infiltrate systems over an extended period to gather intelligence, disrupt operations, or even cause physical damage. The Stuxnet attack, although not specifically targeting solar energy, serves as a stark reminder of how vulnerable industrial control systems can be to such targeted cyberattacks.

Phishing and social engineering attacks also pose a significant risk to solar energy systems. These attacks typically exploit human error, tricking employees or operators into revealing sensitive information, such as passwords or access credentials. Once inside the system, attackers can escalate privileges and gain control over critical components, potentially leading to catastrophic failures or data breaches [13].

Denial-of-service (DoS) attacks, where attackers overwhelm the system with excessive traffic, can disrupt the communication and control functions of solar energy systems. These attacks can lead to temporary loss of control, which, in turn, could cause fluctuations in energy production and distribution. Given the interconnected nature of many solar installations, a successful DoS attack on one system could have cascading effects across the entire network [14].

3. Best Practices for Securing Solar Energy Information Systems

Securing solar energy information systems requires a comprehensive approach involving data protection, robust authentication, access control, and physical and network security measures.

Data Protection: Encrypting data both in transit and at rest is crucial for preventing unauthorized access. Using advanced encryption standards (e.g., AES) ensures the confidentiality and integrity of data transferred between system components, such as sensors, controllers, and storage systems. Regular backups of encrypted data enhance data availability and resilience.

Authentication and Access Control: Implementing multi-factor authentication (MFA) for system access, especially for users with administrative privileges, significantly reduces unauthorized access risks. Role-based access control (RBAC) also limits user permissions based on their roles, following the principle of least privilege. Monitoring user activities helps detect unusual behavior that could indicate security breaches.

Physical and Network Security: In many cases, solar installations are in remote locations, making physical security essential. Using secure, tamper-proof enclosures and surveillance systems helps protect critical components from physical threats. Network security measures, such as firewalls, intrusion detection systems (IDS), and secure wireless protocols (e.g., WPA3), are also necessary to defend against cyber threats and ensure secure data communication within the system.

4. Emerging Technologies and Their Role in Enhancing Security

Emerging technologies like IoT, smart grids, blockchain, and advanced threat detection are transforming the security landscape for solar energy systems.

IoT and Smart Grid Integration: IoT devices, such as sensors and smart meters, allow real-time data collection and improved system monitoring, helping to identify anomalies quickly. While these technologies increase connectivity, they also create new security risks. To address this, IoT devices are equipped with secure boot, encryption, and authentication features, while smart grids use advanced security protocols to monitor and isolate threats within the network.

Blockchain for Data Integrity: Blockchain's decentralized ledger system ensures data integrity by making transactions tamper-proof. It is particularly useful in peer-to-peer energy trading within solar

systems, where secure, immutable transactions reduce fraud risks. Additionally, blockchain can manage device identities, ensuring only authorized devices interact within the network.

Advanced Threat Detection: Machine learning (ML) and artificial intelligence (AI) enhance threat detection by analyzing large data volumes to recognize patterns indicating security breaches. AI-powered intrusion detection systems (IDS) and predictive analytics can forecast vulnerabilities and prevent attacks, providing a proactive approach to system security.

Together, these technologies contribute to a resilient, secure infrastructure by addressing vulnerabilities and enabling real-time responses to emerging threats.

5. Regulatory Frameworks and Industry Standards

The growing role of solar energy systems within the broader energy infrastructure has created an urgent need for dedicated regulatory frameworks and industry standards that address their unique security and resilience requirements. These frameworks and standards offer targeted guidelines to help organizations safeguard solar energy systems against risks, ensure compliance with legal obligations, and strengthen overall security measures specific to solar infrastructure.

Regulatory frameworks for the security of solar energy systems vary by region but generally share common goals: to protect critical infrastructure, ensure the reliability of energy supply, and safeguard sensitive data [15], [16]. In many countries, regulations specific to energy infrastructure set out mandatory requirements for protecting critical energy systems, including those powered by solar energy. These frameworks typically address several key areas, including physical and cyber security, incident response, and disaster recovery. For instance, the European Union's Network and Information Security (NIS) Directive requires member states to adopt national strategies for securing critical infrastructure, including energy systems, and mandates that operators of essential services implement robust security measures and report significant incidents [17]. ISO/IEC 27001, for example, provides a framework for managing the security of information systems, including those in the energy sector, by establishing a systematic approach to managing sensitive company information [18]. In addition to regulatory frameworks, several industry standards have been developed specifically to address the unique security needs of solar energy systems. These standards provide detailed guidelines on how to design, implement, and maintain secure solar energy infrastructures.

For solar energy systems, meeting regulatory standards typically requires a layered approach to security that incorporates both preventive and detective controls. This includes using secure communication protocols tailored to solar infrastructure, performing regular software updates and patches, conducting frequent security audits, and implementing strict access control measures. Adhering to these standards is vital for organizations managing solar energy systems, yet staying compliant can be challenging due to the evolving nature of cyber threats and the technical complexity of modern solar installations.

Keeping pace with regulatory updates is essential, as organizations may need to adapt their security practices and policies frequently to align with new requirements. Regular audits and assessments help identify potential compliance gaps and allow for timely corrective actions. Additionally, fostering a culture of security awareness among employees and stakeholders is key to ensuring that security measures are fully understood and effectively implemented throughout the organization.

As solar energy systems increasingly interact with other parts of the energy grid, there is a growing need for harmonization of standards and regulations across different sectors and jurisdictions [19]. International collaboration and coordination are essential to developing a unified approach to security, particularly as threats become more global in nature.

6. Conclusions

In an era where solar energy systems are becoming essential to sustainable energy, ensuring their security has never been more critical. This paper examined the unique security challenges these systems face, particularly as they integrate into broader, interconnected energy networks. By exploring the vulnerabilities in solar energy infrastructure and the increasing sophistication of cyber threats, we highlighted the pressing need for a multi-layered approach to security.

Key best practices, such as data encryption, authentication, and physical security measures, offer foundational protections, while emerging technologies like IoT, blockchain, and AI-based threat detection add advanced defenses against evolving risks. Additionally, compliance with regulatory frameworks and industry standards helps organizations keep pace with security requirements, supporting the resilience and reliability of solar systems.

Looking ahead, the ongoing security of solar energy systems will depend on continuous adaptation, technological innovation, and collaborative efforts across the industry. As the global shift toward renewable energy accelerates, robust security practices will play an indispensable role in safeguarding this essential infrastructure and ensuring a stable energy future.

References:

- [1] Z. A. Elum and A. S. Momodu, "Climate change mitigation and renewable energy for sustainable development in Nigeria: A discourse approach," *Renewable and Sustainable Energy Reviews*, vol. 76, pp. 72–80, Sep. 2017, doi: 10.1016/j.rser.2017.03.040.
- [2] J. De Boer and C. Zuidema, "Towards an integrated energy landscape," *Proceedings of the Institution of Civil Engineers - Urban Design and Planning*, vol. 168, no. 5, pp. 231–240, Oct. 2015, doi: 10.1680/udap.14.00041.
- [3] H. Ziar, P. Manganiello, O. Isabella, and M. Zeman, "Photovoltaics: intelligent PV-based devices for energy and information applications," *Energy Environ. Sci.*, vol. 14, no. 1, pp. 106–126, 2021, doi: 10.1039/D0EE02491K.
- [4] T. Ahmad and D. Zhang, "Using the internet of things in smart energy systems and networks," *Sustainable Cities and Society*, vol. 68, p. 102783, May 2021, doi: 10.1016/j.scs.2021.102783.
- [5] I. Zografopoulos, N. D. Hatziaargyriou, and C. Konstantinou, "Distributed Energy Resources Cybersecurity Outlook: Vulnerabilities, Attacks, Impacts, and Mitigations," *IEEE Systems Journal*, vol. 17, no. 4, pp. 6695–6709, Dec. 2023, doi: 10.1109/JSYST.2023.3305757.
- [6] A. Azzuni and C. Breyer, "Definitions and dimensions of energy security: a literature review," *WIREs Energy & Environment*, vol. 7, no. 1, p. e268, Jan. 2018, doi: 10.1002/wene.268.
- [7] G. Bedi, G. K. Venayagamoorthy, R. Singh, R. R. Brooks, and K.-C. Wang, "Review of Internet of Things (IoT) in Electric Power and Energy Systems," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 847–870, Apr. 2018, doi: 10.1109/JIOT.2018.2802704.
- [8] A. P. Zhao *et al.*, "Cyber Vulnerabilities of Energy Systems," *IEEE J. Emerg. Sel. Top. Ind. Electron.*, pp. 1–14, 2024, doi: 10.1109/JESTIE.2024.3434350.
- [9] G. Deepa and P. S. Thilagam, "Securing web applications from injection and logic vulnerabilities: Approaches and challenges," *Information and Software Technology*, vol. 74, pp. 160–180, Jun. 2016, doi: 10.1016/j.infsof.2016.02.005.
- [10] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, "A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions," *Electronics*, vol. 12, no. 6, p. 1333, Mar. 2023, doi: 10.3390/electronics12061333.
- [11] J. Martínez and J. M. Durán, "Software Supply Chain Attacks, a Threat to Global Cybersecurity: SolarWinds' Case Study," *IJSSE*, vol. 11, no. 5, pp. 537–545, Oct. 2021, doi: 10.18280/ijssse.110505.
- [12] D. P. F. Möller, "Ransomware Attacks and Scenarios: Cost Factors and Loss of Reputation," in *Guide to Cybersecurity in Digital Transformation*, vol. 103, in *Advances in Information Security*, vol. 103, Cham: Springer Nature Switzerland, 2023, pp. 273–303. doi: 10.1007/978-3-031-26845-8_6.
- [13] S. Tufail, I. Parvez, S. Batool, and A. Sarwat, "A Survey on Cybersecurity Challenges, Detection, and Mitigation Techniques for the Smart Grid," *Energies*, vol. 14, no. 18, p. 5894, Sep. 2021, doi: 10.3390/en14185894.
- [14] M. N. U. Islam, A. Fahmin, Md. S. Hossain, and M. Atiquzzaman, "Denial-of-Service Attacks on Wireless Sensor Network and Defense Techniques," *Wireless Pers Commun*, vol. 116, no. 3, pp. 1993–2021, Feb. 2021, doi: 10.1007/s11277-020-07776-3.
- [15] S. Abdelkader *et al.*, "Securing modern power systems: Implementing comprehensive strategies to enhance resilience and reliability against cyber-attacks," *Results in Engineering*, vol. 23, p. 102647, Sep. 2024, doi: 10.1016/j.rineng.2024.102647.

- [16]S. Dong, J. Cao, and Z. Fan, "A Review on Cybersecurity in Smart Local Energy Systems: Requirements, Challenges, and Standards," 2021, *arXiv*. doi: 10.48550/ARXIV.2108.08089.
- [17]J. D. Michels and I. Walden, "Cybersecurity, Cloud and Critical Infrastructure," *SSRN Journal*, 2021, doi: 10.2139/ssrn.4204847.
- [18]F. Kitsios, E. Chatzidimitriou, and M. Kamariotou, "The ISO/IEC 27001 Information Security Management Standard: How to Extract Value from Data in the IT Sector," *Sustainability*, vol. 15, no. 7, p. 5828, Mar. 2023, doi: 10.3390/su15075828.
- [19]A. Gopstein, C. Nguyen, D. S. Byrnett, K. Worthington, and C. Villarreal, "Framework and roadmap for smart grid interoperability standards regional roundtables summary report," National Institute of Standards and Technology, Gaithersburg, MD, NIST IR 8284, Jan. 2020. doi: 10.6028/NIST.IR.8284.

Could Facebook's user interface be more usable for seniors?

Dragana Bodiřoga^{1,2} and Dragan Ivetić¹

¹ Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia

² QCerris, Suvoborska 29, Belgrade 11000, Serbia

draganamihajlovic97@uns.ac.rs; ivetic@uns.ac.rs

Abstract:

Today, a significant number of seniors (aged 55+) use the Internet and social media, yet these technologies often remain inadequately adapted to this population. As many as 282 million users of the social network Facebook are over 55 years old, which is more than the fourth most populous country on the planet, Indonesia. Difficulty finding settings, insufficiently clear feedback, frequent updates on the user interface and poorly implemented help features are some of the shortcomings of Facebook's interface in terms of adaptability to seniors. This paper will present the user interface of the social network Facebook in terms of adaptability for seniors. Additionally, ways in which these shortcomings can be eliminated in order to make the user interface of the social network Facebook more usable for the seniors are presented.

Keywords:

seniors (55+), usability, Facebook

1. Introduction

Population aging has entered an accelerated phase in most countries, rising widespread concern globally [1]. Social isolation and loneliness are increasingly recognized as significant public health and policy challenges across all age groups. Over the course of the UN Decade of Healthy Ageing (2021-2030), the Demographic Change and Healthy Ageing Unit will address social isolation and loneliness as key themes intersecting with the four main action areas of the Decade [2]. Seniors, whether by choice or necessity, must adapt to the new digital age and the rapidly evolving information technologies that accompany it.

In this context, Facebook is the most widely used social media platform globally. It has connected people across the world, enabling them to stay in touch with friends and family regardless of distance. Facebook has also become an important platform for social activism and political discussion, as well as a good tool for disseminating news and information [3]. As of the first quarter of 2024, the global population was approximately 8.1 billion [4], with around 5.44 billion Internet users and around 5.07 billion social media users [5].

These are indicators of a large share of the world's population in modern information and communication technologies, a trend that continues to grow. The share of Internet users worldwide over the age of 55 is approximately 10.2% [5]. Furthermore, the number of Facebook users in the first quarter of 2023 over the age of 55 was 282 million, which is around 12.5% of the platform's total audience (Figure 1) [6].

Due to the large number of elderly people on this social network, more and more often, the question arises whether the user interface of this social network is adapted to this group of users.

Numerous studies indicate that many cognitive functions such as episodic memory, attention, executive functions, spatial orientation, visual perception, information processing speed and verbal memory, decline with age [7]. These abilities are crucial factors that limit older people from effectively using the Internet and social media.

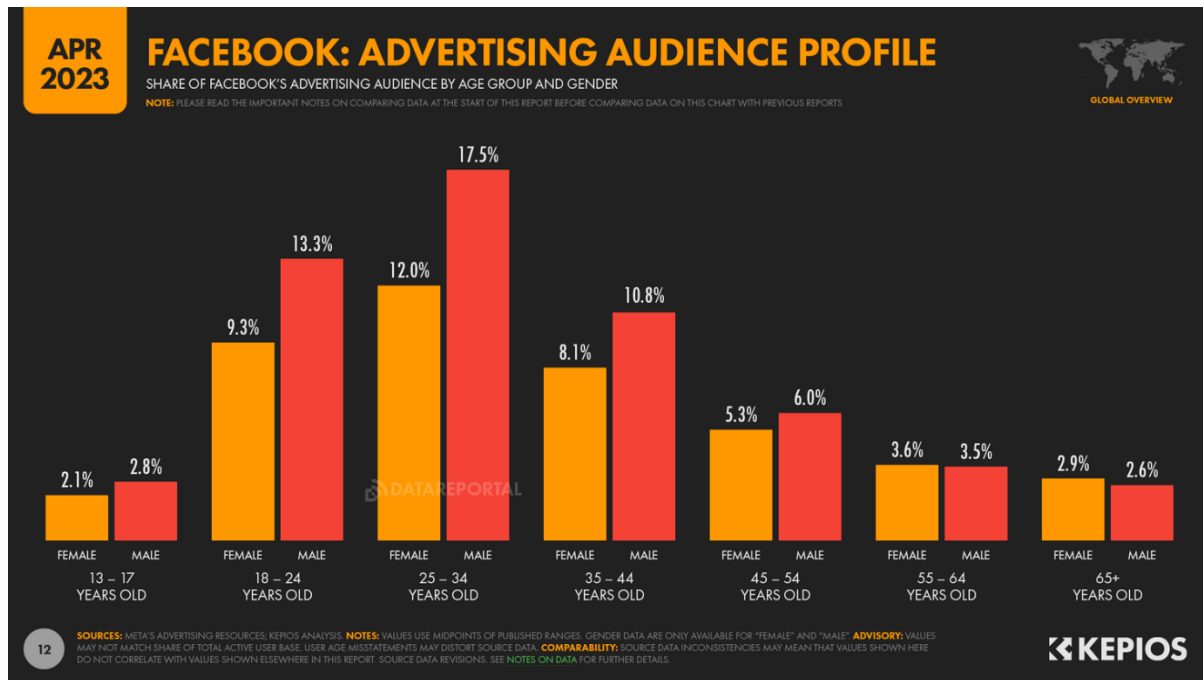


Figure 1: Share of Facebook users by age and gender [6]

Limitations like blindness, despite the availability of assistive technology, may even prevent the use of online social networks if no further additional assistance is available [8]. Short-term memory decline presents a greater challenge in user interface (UI) design for older people. Additionally, physical and motor impairments become more pronounced with age. An example of this is a tremor that can significantly affect navigation through the interface, particularly on touch devices. While older people can perform certain actions just like younger individuals do, it often takes them longer. They may easily lose focus, so care should be taken to avoid requiring multiple actions on a single display. Combining actions can lead to confusion, frustration, and ultimately, abandonment of a task or webpage.

Research involving focus groups of participants aged 65 to 87 identified various barriers to Internet use among older adults. These barriers include intrapersonal barriers ("I'm too old for that"), functional (related to memory decline and spatial orientation), structural ("it costs too much to own a computer") and interpersonal barriers ("no one teaches me how") [9].

2. Research methodology

This paper deals with the analysis of the user interface of the social network Facebook in terms of use by seniors (55+) taking into account different sources of literature. The goal of this research is to show how well the user interface of the social network Facebook is adapted to the elderly population, given that as many as 12.5% of Facebook users are over 55 years old [6].

This is also a form of theoretical research in which the results of previous studies are taken. This research identifies and proposes solutions for relevant problems of the interface of the social network Facebook by better understanding the needs of end users.

By studying already existing literature, reading sites and forums for seniors such as Senior net and My Senior Portal, interviewing people from the environment over the age of 55, using existing proposals for adaptation user interface for seniors, as well as guidelines for the development of the user interface, this study analyzed what are the risks of seniors not using social network Facebook and how the user interface can improve. Also, this paper includes what people did earlier and what they reported as a lack of Facebook. The features that seniors use the most were observed and analyzed, and based on that, improvements were proposed.

3. Previous analyzes of the state of the user interface of the social network Facebook for seniors

Adapting the UI of social media platforms, including Facebook, to accommodate users across all ages is a significant challenge. The action must be presented in a simple way so that everyone can perform it intuitively and with ease.

There are three main problem areas when designing a UI: utilizing screen space, interaction mechanisms and the overall design [10]. Among the most important things related to the UI is social media functionality, including interface and navigation - 46%. Additionally, the aesthetics of a social media application are important to users. This includes consistent branding, colors, high resolution images, clear and sharp buttons, appropriate text size and fonts [11].

Numerous studies highlight that many seniors individuals continue to face mental health challenges. As joining virtual social communities has shown to improve the mental health among older people, the UI of these websites needs to be adapted to them. A study focusing on the adoption of social networks to older adults, specifically Facebook, provided several recommendations for enhancing its interface design [12]. The study involved participants with an average age of 72, ranging from 67 to 81 years old. Based on the survey conducted, privacy was the main reason for giving up the use of the social network Facebook by older people, and it was proposed to modify some privacy options on this social network, such as adding "Publish to" and "Show from" actions. The participants of this study reacted positively to the features that provided information about the family, and suggested that the interaction family should be more family-oriented. Also, the respondents gave more importance to photos than to other types of content. They also suggested filtering media content. Finally, the respondents emphasized the need for a simple interface and avoidance of terminology that is not in the user's native language.

On the other hand, there is a simplified version of Facebook that is only about 2 MB in size (in comparison, the Facebook application for Android is 51 MB) - Facebook Lite. It was launched in 2015 designed to work in areas with slower or limited internet connections [13]. Facebook has more animations and design displays. Facebook's interface consumes more mobile data as compared to the minimalistic interface offered by Facebook Lite. Although the Facebook Lite interface has been simplified, unfortunately the problems that exist on the Facebook interface have not been overcome. Below is a table (Table 1) highlighting application design differences between Facebook and Facebook Lite [14].

Table 1:
User interface differences between Facebook and Facebook Lite [14]

Feature	Facebook	Facebook Lite
Design	Modern, purple-blue scheme	Simple, white background
Layout	Tabs based interface	Icon-based interface
Color Scheme	Gradient approach with darker hues for the top bar	Solid graphics lighter hues

In an experiment [15] conducted with people over 60 years, including both those with technical education and those without, participants were asked to perform the following tasks on Facebook: view notifications, share a post, like, comment, change profile picture, change cover photo, send a friend request, tag a friend, post and send a message. Some of the respondents had a problem with the visibility of the notification icon, quite a few of them did not know where to click to share a post, most of them had no problems liking a post, and a similar situation was with commenting on a post. As for changing the profile photo, divided mixed results were presented. While some participants did not have any problems, while others had problems due to the visibility of the button for changing the photo. The situation is similar to changing the cover photo. Sending a friend request proved challenging for seniors who did not know how to send a request if the person was not currently on their homepage. Tagging friends was problematic because many were unfamiliar with the necessity of using "@" before a friend's name. Additionally, the post button only appears after users begin typing in the post text field, which confuses some participants. The participants had no problems in sending messages.

4. A solution proposal for improving the user interface of social network Facebook

Considering that older people use mobile devices more than desktop devices, the focus of this study will be on the UI of Facebook's mobile application.

4.1. Audio presentation

The social network Facebook offers customizable audio notifications. In the settings, it is possible to select the actions for which the user would like to receive notifications, such as: comments, tags, reminders, friend requests, birthdays, groups, events, etc. By default, the sound for all notifications is a generic "drop" sound. What is good is that there is a possibility to select the desired sound for each notification category separately. However, configuring this on a mobile device can be challenging, especially for the seniors. For this demographic, customized audio notifications can be particularly beneficial, since they can know what category the notification is without getting up and picking up the phone. Additionally, Facebook provides audible feedback for actions like posting a comment, which further enhances the user experience.

By interviewing, one of the respondents (age 67) said that one of the disadvantages is the volume of notifications. Seniors are less likely to turn up the volume. One of the ways to regulate and improve it is to set the volume based on the date of birth (age of the user). This would not force the seniors and would improve their experience on this social network.

4.2. Video presentation

The homepage can be divided into two main parts (Figure 2).

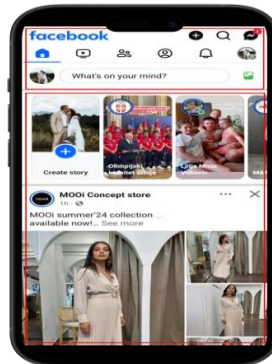


Figure 2: The homepage of the social network Facebook on the mobile application

The first part consists of three sub-levels:

- Space for the logo, a shortcut to add a new post, a shortcut for searching, and a shortcut to open the Messenger application;
- Tabs for homepage, Reels, friend requests, personal profile, notifications and settings;
- Field for adding a new post.

The second part is the space for adding and viewing stories, as well as the home page, which occupies the central part of this view.

The first part satisfies Nielsen's principle: match between the system and the real world. Icons such as friend requests and notifications are intuitive and clear and should not cause any discomfort for the seniors. However, a disadvantage mentioned by the interviewee in the survey (age 59) is the field where further navigation would lead to settings. He said that it's unintuitive as it is represented by the user's profile picture. This is a disadvantage in terms of interface especially with inexperienced older users who would like to enjoy the use of this social network.

One of the actions that generates satisfaction among older Facebook users is posting photos [12], and this action will be presented below. Clicking on the field "What's on your mind" (Figure 3) opens a new view in which the possible options are presented. Selecting the "Photo/Video" field (Figure 3) allows the users to access the phone's gallery, where the desired photo or video can be selected.

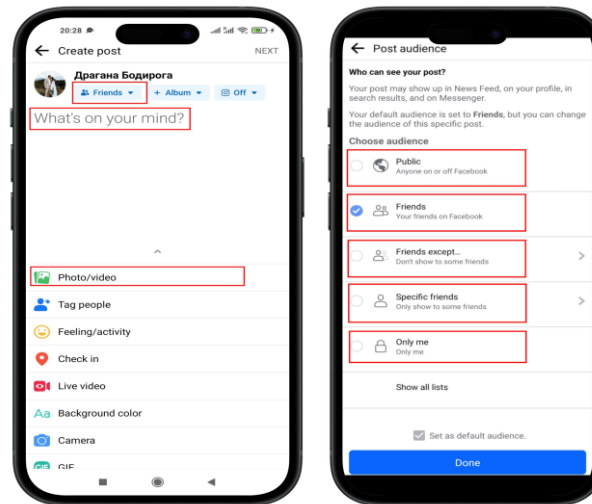


Figure 3: Adding a new post and setting post visibility

Before publishing the photo(s), the user can choose who can see the photo(s) (Figure 3). Although this can be set in the settings, the user is given the option to decide on the visibility of their post directly before posting it. However, the question is whether this is intuitive enough for an older user and whether they will pay attention to this feature on a given display.

The default visibility of posts is set to "Friends" and hides the posts from users who are not in the friends list, which is good in terms of user privacy (and is also required by most data privacy laws across the world, including the European Union's GDPR). In one study conducted [12], several participants, whose average age was 72 years, were unaware of the existence of this feature at this part of the UI, indicating that its usability could be improved. On the other hand, two interviewees stated that this is a good default setting because they were not even aware of the possibility of these settings.

When selecting media, users can choose multiple photos or videos by tapping the "Select Multiple" button (Figure 4). Once activated, the button turns blue, and each media item in the gallery is marked with a circle in the upper right corner displaying a serial number. This feature provides clear feedback to users, particularly benefiting older adults who prefer to have ongoing control and visible status updates.

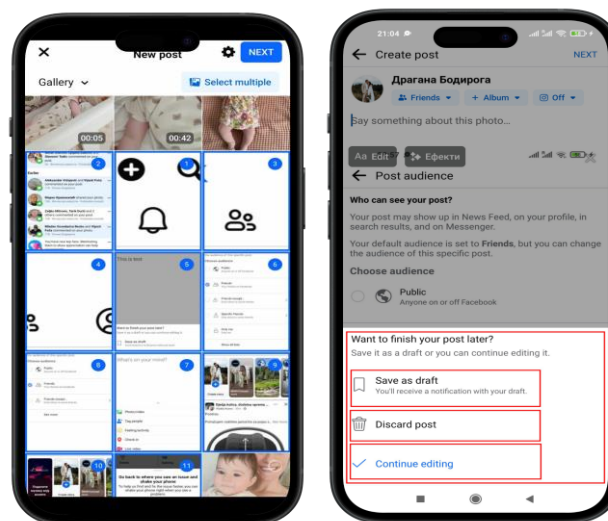


Figure 4: Multiple selection of photos/videos and action confirmation message

If the user accidentally navigates backwards, a question appears at the bottom of the screen in which he can confirm whether he wants to exit or save the draft post for later (Figure 4). This part is maybe best designed and adapted to seniors (aged 55+). The space that should no longer be the focus of the user is slightly dimmed, the message about leaving or continuing the action is highlighted and the user's attention is directed towards it. The elderly have a weaker focus, so this is a good way not to disturb their attention and not to have the user interface take them to another page.

One of the disadvantages of the UI of the social network Facebook, mentioned by the interviewed user (aged 73), is the smaller number of clearly highlighted feedback actions. For example, when commenting, older users may not pay attention to the progress of posting a comment, which is shown as a label under the comment with text that is slightly transparent (Figure 5). Feedback for the just published comment is presented according to the same principle (Figure 5).

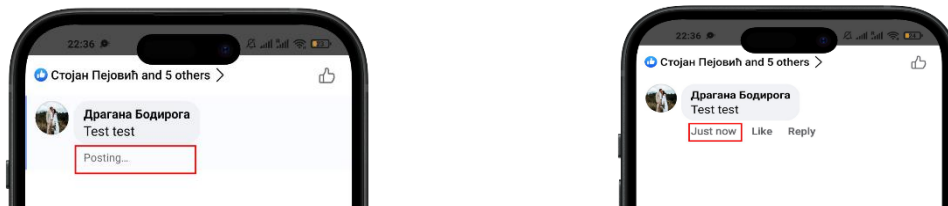


Figure 5: Feedback of the progress of commenting on the post and the successfully posted comment

An example of a well-visible system status is the navigation bar (Figure 6), which clearly shows the page that is currently presented.

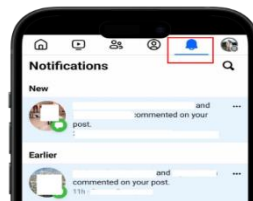


Figure 6: Navigating through the main menu and a good example of system status visibility

A major challenge in terms of the adaptation of older people to this social network, is the frequent change of versions that are published in production. As all users, especially older ones, want consistency and standards when using an application, these frequent updates in which the UI changes can hinder user retention.

Additionally, while users can report problems through the "Help and Support" section, the process is not intuitive. Users are instructed to shake their phone on the page where the problem occurs to report it (Figure 7). This method has several drawbacks, the most significant being that seniors may struggle to access the specific page where the issue happened or may have difficulty performing the shaking action. It is important to avoid actions that require large gestures or movements, such as shaking, as these can be challenging for many users, particularly older adults.

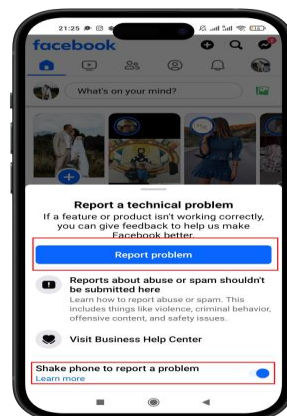


Figure 7: Report a problem by shaking the phone

5. Conclusions

Social media usage among 55-year-olds continues to grow. The number of social media users is large, but it is necessary to offer a way to integrate social media and traditional communication. This especially applies to older people who have difficulty accepting new information technologies, loneliness, alienation and social isolation. They need to be offered a solution that would allow them to find interests, friends and activities that they can do in the traditional way they are used to. The prerequisites for the seniors to be able to use social media are, first, access to the Internet, a device, as well as knowledge about information and communication technologies and social media which are necessary to carry out actions without difficulty and without causing frustration or discomfort.

The observed problems are not unique to the social network Facebook, but this social network served as an example to illustrate the problems that seniors (aged 55+) face with social network UI, especially since Facebook is the most used social network by 55+ users.

The user interface of the social network Facebook could be more usable for older users if the previously mentioned shortcomings are removed in a way that would allow seniors to use certain functions more easily.

The most problematic disadvantage stated by the interviewed respondents is reporting the problem. One of the interviewed said: "I will never be able to report any problem".

Returning to the place of the problem and shaking the phone hard is challenging for older users, especially if we take into account the physical capabilities of older people. Also, since mobile phones are devices that are used most often by the seniors, this can also result in the activation of some functions located on the sides of the phone during shaking: volume up/down, turning off the phone, locking the phone, etc. These functions are very important for the population older than 55 years, especially for those who are new to this application. A possible solution is to determine the place on the interface where the help button would be, and based on the positioning on the interface, the user would be guided how to solve the problem.

Feedback about success/failure should be much more and must be more visible. A better solution would be pop-up windows for registered users over 55 years old.

The text on the interface should be localized and fewer terms that are not in the user's native language need to be displayed. New information technologies can help seniors with health problems, but we should also look at those who are in good health but have excess free time, are lonely, more socially isolated and are ready to get involved in some activities that would improve their everyday life and this will be the focus of further research. Perhaps the older population wants their own social network with content and features adapted to their interests and needs. The new social network would have a UI adapted to this age category. Another possibility is to create two interfaces for older and younger users, in which the needs of both age groups would be fully met.

UI designers cannot successfully design an application if they do not understand the needs and wishes of the users. It is important to focus on each feature and clearly define what function it offers to a user. Currently, the social networks that support the largest number of users do not fully meet the needs of the seniors. The interface must be simple and consistent, but the Facebook interface is frequently updated and this negatively affects the experience of older people. Additionally, it is necessary to provide safe points at all times and it is necessary to give the user the option to return to the homepage. That is why these changes should have a minimal impact on the UI, and adding new features should not change the UI much.

The redesign of the Facebook social network, which would be adapted to older people, would increase the number of older people on this social network. Activities that are easier for users to perform as well as those that are more difficult for them to perform should be checked. Based on this, it would be determined in what way and which parts of the interface should be changed.

In conclusion, this paper underscores the imperative for the designers of the interface of social networks, especially the social network Facebook as the most popular social network among seniors, to make changes to certain parts of the interface so that seniors can more easily accept, use and enjoy all the features offered by this application. Proposed ways to solve some of the shortcomings of the user interface in this paper will help designers of social networks Facebook to create a more usable interface for the elderly.

Acknowledgment:

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad” (No. 01-3394/1).

References:

- [1] Tao Hea, Changqin Huang, Ming Lib, Yuqiong Zhouc, Shihua Li: Social participation of the elderly in China: The roles of conventional media, digital access and social media engagement, vol. 48, pp. 1-3, 18 January 2020.
- [2] World Health Organization - Social isolation and loneliness. URL: <https://www.who.int/teams/social-determinants-of-health/demographic-change-and-healthy-ageing/social-isolation-and-loneliness>, date accessed: 12 October 2024.
- [3] Mayorie María Vargas Vega: The Impact of Facebook on Society: A Look into its History and Significance. URL: <https://medium.com/@mayoriemaria/the-impact-of-facebook-on-society-a-look-into-its-history-a-nd-significance-3982526eaacd>, date accessed: 12 October 2024.
- [4] Worldometer. URL: https://www.worldometers.info/world-population/#google_vignette, date accessed: 15 August 2024.
- [5] Ani Petrosyan, Age distribution of internet users worldwide 2024. URL: <https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide>, date accessed: 10 August 2024.
- [6] Datareportal: Facebook users, stats, data and trends. URL: <https://datareportal.com/essential-facebook-stats>, date accessed: 5 August 2024.
- [7] Paulina Zadro, Sanja Šimleša, Marina Olujić, Jelena Kuvač Kraljević: Promjene kognitivnih funkcija u odrasloj dobi, URL: <https://hrcak.srce.hr/file/253827>, date accessed: 3 August 2024.
- [8] Anja K. Leist: Social Media Use of Older Adults: A Mini-Review, national library of medicine, Gerontology 2013, Vol. 59, No. 4. doi: <https://doi.org/10.1159/000346818>.
- [9] Lee B, Chen Y, Hewitt L: Age differences in constraints encountered by seniors in their use of computers and the Internet, Computers in Human Behavior Volume 27, Issue 3, May 2011, pp. 1231-1237 URL: <https://www.sciencedirect.com/science/article/pii/S0747563211000070?via%3Dihub>.
- [10] Erik G. Nilsson: Design patterns for user interface for mobile applications, Advances in Engineering Software Volume 40, Issue 12, December 2009, Pages 1318-1328. doi: <https://doi.org/10.1016/j.advengsoft.2009.01.017>.
- [11] Jackie R. Hayes: User Interface Design for Online Social Media, March 2014, URL: <https://digitalcommons.calpoly.edu/grcsp/108>.
- [12] Gonçalo Gomes, Carlos Duarte, José Baptista Coelho, Eduardo Matos: Designing a Facebook Interface for Senior Users, Scientific World Journal v.2014; 2014, 2 February 2014, URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3929575/#>.
- [13] Tzach Hadar, Dekel Naar, Made in Tel Aviv: Facebook Lite, 19 August 2020, URL: <https://tech.facebook.com/engineering/2020/8/facebook-lite-made-in-tel-aviv/>, date accessed: 15 August 2024.
- [14] Rajat Garg: Facebook vs. Facebook Lite – All the Differences Explained. URL: <https://www.socialappshq.com/facebook/fb-vs-fb-lite/>, date accessed: 30 August 2024.
- [15] Rizwana Saeed, Zaibunnisa khan, Khizra Naqui, Sumaira Iqbal and Fizza Aftab: Usability Issues on Facebook for Older Adults, MEJSR Volume 22 Number (7), 2014, URL: [https://idosi.org/mejsr/mejsr22\(7\)14/22.pdf](https://idosi.org/mejsr/mejsr22(7)14/22.pdf).

Design and Development of Laser Engraving Machine

Evgenija Divitarova¹, Andrijana Bocevska²

^{1,2} University St Kliment Ohridski, IMaj nn, 7000 Bitola, R. North Macedonia

divitarova.evgenija@uklo.edu.mk; andrijana.bocevska@uklo.edu.mk

Abstract:

Hands-on laboratory instructions are essential to improve the student learning feature in engineering education. In the area of Computer Aided Design (CAD) and Computer Aided Manufacturing (CAM), Computer Numerical Control (CNC) machines with their significant cost limit the accessibility of machines to the students. In this paper, a low-cost, Build-Your-Own (BYO) two axis (X and Y) vertical Computer Numerical Control (CNC) prototype is developed using off-the-shelf components, stepper motors with drivers, Arduino open-source microcontroller, and open-source motor control software. The machine provides better access and insights to the students for easy operation as well as the capability to do various sophisticated precision engraving, whether for practical or artistic purposes.

Keywords: Laser, CNC, LaserGRBL, engraver, G-code, Arduino

1. Introduction

CNC engraving is a process where computer-controlled machine tools etch or carve designs, letters, or patterns into a surface. CNC stands for Computer Numerical Control, which refers to use a computer to control the movement and operation of the machine. In the case of CNC engraving, a design is created in a software program. This design is then translated into a language (commonly G-code) that the CNC engraving machine can understand. This code directs the machine where to move and how fast to go, among other things. It's a subtractive process where the machine uses rotating cutters, or burrs, to cut away material to create the desired shape. CNC engravings are of two types: CNC milling engraving and CNC router engraving [1].

Laser engraving can be considered a type of CNC process where a CNC laser engraver uses a laser beam to create text or designs on various materials. It's a non-contact process that uses a powerful laser beam to cut or etch a design into a material. Laser engraving is a popular method for creating intricate designs on metal, plastic, wood, and other materials. The light beam used in the process is highly concentrated and precise so that the design can be detailed and intricate.

The CNC laser engraver is the main tool used in the process. It contains a computer-controlled laser module/head, the source of the focused light beam. The laser engraving process is fairly straightforward. First, the material to be engraved is placed into the laser engraving machine. Then, the computer-controlled laser module traces the design onto the surface, burning away material from the surface to create the desired design. In addition to its accuracy, laser engraving is a fast and economical way to add a personalized touch to items.

The goal is to come up with an in-house solution or create your own, Do It Yourself (DIY) design, development, and testing of laboratory equipment and systems. Therefore, low-cost, desktop and small-to-medium sized CNC machines such as mills, lathes, routers, and laser and plasma cutters can be built with an open structure so that the construction of the machine is clear and understandable to the students. The cost of the machines may be minimized by using inexpensive yet sufficiently powerful off-the-shelf stepper motors and servomotors with motor drivers. Legacy controllers can be replaced with PC-based control systems with low-cost data acquisition cards and sensors.

Lasers can have different types based on the gain medium they use. This gain medium also dictates the optical power of the laser. Commercially, the following three types of laser engravers are common [2]:

- Fiber Laser Engravers - are the type of solid-state laser with fiber optic as the gain medium.

They have a wavelength between 780 nm and 2200 nm, which is readily absorbed by most materials, even metals. They are the most powerful laser engravers on the market and are commonly used in metal engraving needs.

- CO2 Laser Engravers - have optical power which can be as low as 40W and go up to hundreds of watts. Their wavelength being close to 10,600 nm is absorbed by most materials but reflected by metals. They are not the preferred choice for metals but excel at non-metals engraving and cutting.
- Diode Laser Engravers - one of the lasers that is commonly used in desktop and hobby laser engravers. They may not be as powerful as CO2 or fiber lasers, yet they are compatible with a wide range of materials for engraving. As another benefit it can be mentioned that they are largely less priced compared to the other two lasers.

The fundamental goal of this paper is to discuss the process of designing and developing a low-cost, Build-Your-Own (BYO) two axis (X and Y) vertical laser engraving machine which is convenient to control by the Arduino CNC. The prototype is developed using off-the-shelf components, stepper motors with drivers, Arduino open-source microcontroller, and open-source motor control software and can be used to engrave designs on wood. The rest of the paper is structured as follows. Next section discusses related works, whereas the design on the laser engraving machine and the adjustment of the engraving parameters are described in Section 3 and Section 4, respectively. Section 5 offers an overview of the obtained results, and we provide concluding remarks in the last section.

2. Previous work

A technique used to develop a low-cost CNC laser engraving machine capable of engraving 2D images on materials such as plastic, wood, paper, etc. is proposed by [3]. They used the Benbox software to provide images and control the hardware's operation. The research done in [4] has focused on the design and fabrication of CNC engraving machines based on open-source electronics, with an emphasis on cost-effective solutions and specific applications such as wood engraving and milling. The importance of the machines lies in their ability to provide efficient and precise engraving solutions for a variety of materials, reducing errors and improving speed compared to manual engraving methods. The aim of [5] is to design a compact, cheap and low power CNC Laser Engraver that is also easy to operate in order to reduce complexity, cost and manual work. Their laser engraver uses the open source LaserGRBL software and can engrave onto surfaces like wood, Acrylic or plastic. The paper [6] discusses the design and implementation of two-dimensional CNC router which can engrave 2D & Gray scaled images or pictures with help of high watt burning laser module on surface which can be a paper, wood, leather, plastic or foam. A technique used to design and develop a CNC Laser Engraver and 2D Plotter (all-in-one machine) is proposed in the paper [7]. The specialty of this machine is that the user can change the tool easily whenever they want to perform many operations such as engraving an object and can also produce a 2D drawing for a particular object in A5 size sheet. Both laser engraver and 2D Plotter respond to G Codes which can be generated by the software. Research project presented in [8] is carried out for the purpose of fulfilling the following two objectives: Design and mechanical fabrication of a laser engraving machine, following the 2-axis structure; and Building program to control the engraving machine according to the design drawings. The machine can work on many different materials such as mica, wood, cloth, fruit peel, etc. The study [9] addresses issues in CNC operator training and shows how the open-architecture control system may be used in practice on a homemade 3-axis mini-CNC machine. The CNC machine tool described in the article is intended mainly for instructional purposes, it is inexpensive to build and requires just a single computer that is compatible with CNC software. A low-cost laser engraving machine to engrave plastics, acrylic, glass, wood, cardboard, and leather is the aim of [10]. The authors have found that this laser engraving process has a higher precision and accuracy as compared to traditional embellishing and embossing. CREO 2.0 software is used for the 3D modeling and simulation of the machine. Arduino and different controller boards are tried and used in assembling the machine in less time. The final assembly is experimentally verified which is based on a 2D gantry that is mounted with a laser LED payload for laser engraving on various compatible materials.

3. Design of the Laser Engraving Machine

Over the last decade, we are witnessing significant advances in laser engraving technology. Consequently, woodworking, furniture production and craft industries have seen an increase in the significance of CNC wooden laser engravers. The components of a laser engraving machine are shown in Figure 1. This machine uses a power-laser beam of 1.5W to engrave the material controlled by computer numeric instructions. The machine can be applied for photo engraving, letter cutout, name and logo personalization. This machine is designed to make the laser engrave in wood having an Arduino and the LaserGRBL software as machine automation basis. Arduino is today's most popular open-source hardware prototyping platform. Arduino receives the G-Code over serial port from the G-code interpreter and sender utility tool running on PC and gives the actuation commands to the stepper motor controller. The G-code program is loaded into the Arduino microcontroller.



Figure 1: Component of the Laser engraving machine

The technique uses the LaserGRBL software to make the process easy [11]. LaserGRBL software is a popular choice for controlling CNC laser engravers due to its versatility and user-friendly interface. It provides features such as file importing, image editing, laser power control, speed settings, and more, all available via a user-friendly interface. Users can often modify settings, adjust parameters, and enhance the software to add new functionality.

This 2D engraving machine utilizes two stepper motors for the X and Y axes (since the laser focus is manually adjusted it was not necessary to set Z). This CNC laser engraver machine is used to engrave desired designs on wood. First, it is necessary to choose an image from LaserGRBL software or design your pattern in software. Arduino passes the signal from the software to the CNC shield. CNC shield operates the drivers as per the signals and moves the motor as per the requirements. The constant laser beam engraves the material, with adjustable intensity controlled by software. Here, software does all the work such as tracing the vector image on board, maintaining the speed of drivers, etc. Hardware model of CNC laser engraver is shown in Figure 2.



Figure 2: Hardware model of CNC laser engraver

4. Adjusting the parameters for engraving

Depending on what we want to be engraved or cut, adjustments must be made to the engraving parameters, Figure 3. If we are engraving an image, it is necessary to transform it into black and white. For this purpose, LaserGRBL software offers several options. According to the needs of the client, Brightness, Contrast, White Clip can be easily adjusted. There are three options for Conversion Tool (Line to Line Tracing, 1bit BW Dithering, Vectorize). We then choose the direction of engraving the image (Horizontally, Vertically or Diagonal).

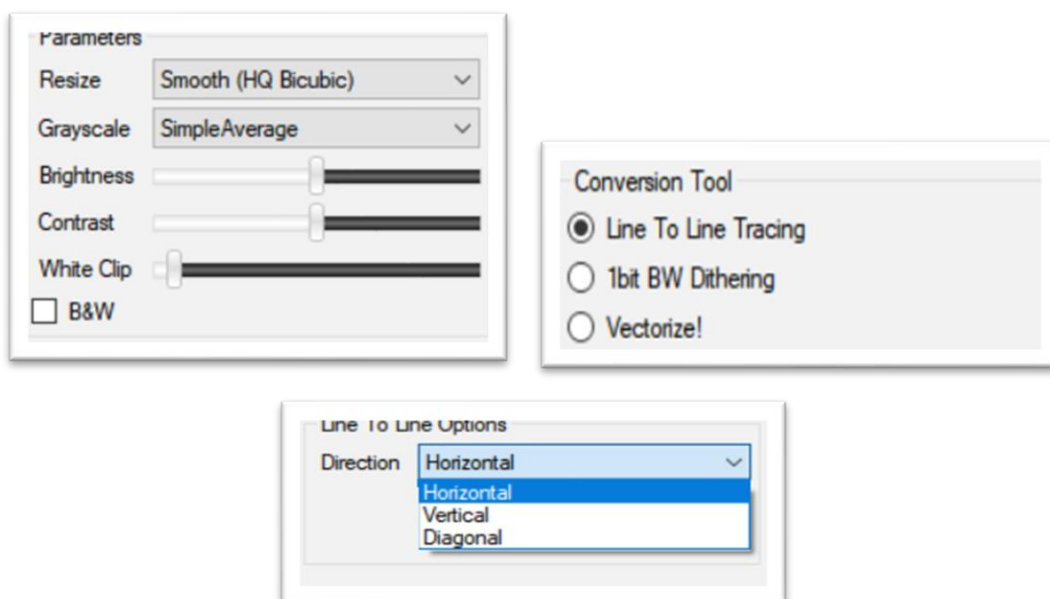


Figure 3: Adjusting the parameters for engraving

Before using a laser engraving machine, it is crucial to test three major controls in the window Target image that influence how the engraving looks: lines per mm, engraving speed, and power level, Figure 4. Properly setting this controls ensures the depth, clarity, and efficiency of the engraving and cutting. Setting the speed and power too high or too low can result in a shallow engraving, blurry results, or even burning the material.

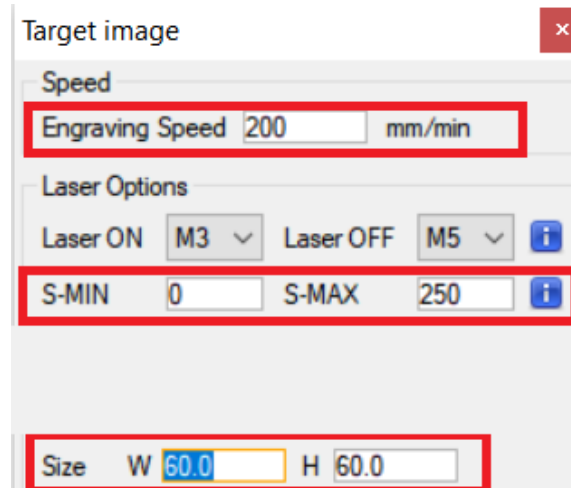


Figure 4: Adjusting the controls in the Target image window

The most important part of a laser engraver is the laser. For the laser to achieve the ultimate precision and accuracy, the focal length must be adjusted to the thickness of the material you are engraving. Proper focusing ensures the energy of the laser beam is concentrated on one point resulting in an accurate high-quality effect. Improper focusing will result in unclear, uneven, or no effect at all.

After the image is selected and the laser engraving settings are configured, the engraving process begins. The software informs us of the engraving completion time, which is determined by the material, power settings, and design complexity. After the start of the engraving process, LaserGRBL software allows the progress to be followed and stops it on an emergency basis if needed.

5. Results

In this section, the results obtained through engraving onto the surfaces are observed. The original images are shown on the left, and how they were engraved on the right on Figure 5. The outcome acquired is a wonderfully engraved complex and an exact picture that is modified to the requirements and fulfilment of the user. The surfaces used are wood, as a solid, natural material known for its durability and unique aesthetic appeal. Different types of wood can have very different characteristics when it comes to laser engraving. The results obtained are exact same to the design provided in LaserGRBL software. While it may not be as straightforward as traditional printing, the advent of laser engravers has made this process much more accessible. This type of small machines reduces the problem of maintenance and can be easily installed at a low cost.

From the results we can conclude that laser engraving stands out as a superior method for transferring images to wood when compared to alternative methods due to its precision, durability, consistency, and versatility. It allows for the creation of highly detailed and permanent images that are resistant to wear and environmental factors. Additionally, laser engraving is clean, safe, and efficient, offering significant control over the engraving process without the need for direct contact with the material, thereby minimizing the risk of damage. While the choice of image transfer method may depend on project specifics and personal preferences, laser engraving is often preferred for its high-quality results and ability to replicate designs accurately across multiple pieces.

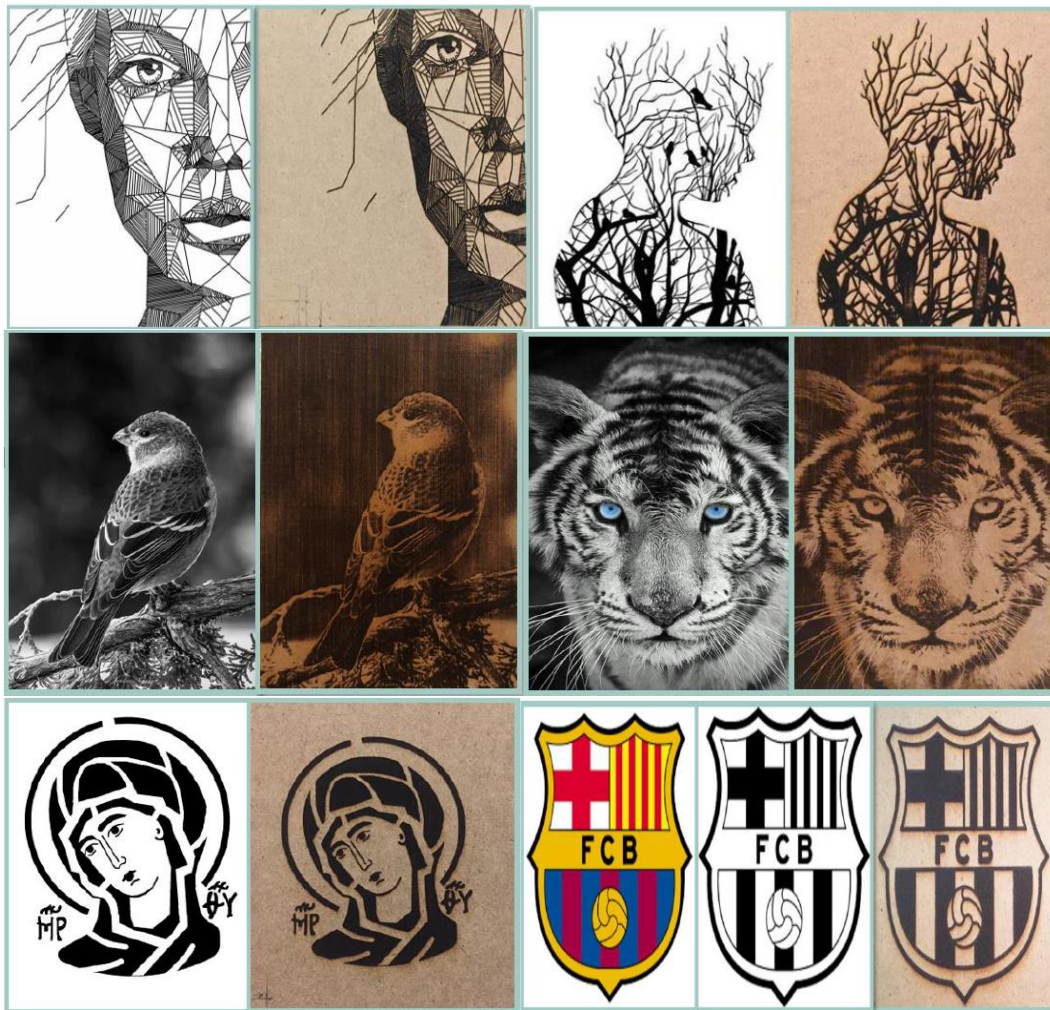


Figure 5: The original images are shown on the left, and how they were engraved on the right

6. Conclusions

This paper discussed the design and development of a laser engraving machine which is convenient to control by the Arduino CNC to engrave desired designs on wood. We can start by choosing an image from Benbox software or by designing a custom pattern in software. Arduino passes the signal from the software to the CNC shield. The CNC shield operates the drivers as per the signals and moves the motor as per the requirements. The constant laser beam engraves the material, with adjustable intensity controlled by software. Software does all the work such as tracing the vector image on board, maintaining the speed of drivers, etc. It is accessible and perfect utilization for small and medium scale industries. The results indicate that laser engraving is a superior method for transferring images onto wood compared to other techniques, thanks to its precision, durability, consistency, and versatility. Using this type of CNC laser engraving machine, users are able to add personalized touches such as logos, photos, and artwork. By offering greater personalized experiences, a wider customer and boost revenue can be attracted. The combination of a compact design, plug-and-play features, clear instructions, user-friendly software, and technical support makes CNC laser engravers relatively easy to install for users at different skill levels. The precision of CNC laser engraving machines ensures that designs are accurately engraved onto materials, resulting in high-quality finished products. Also, the automated nature of CNC laser engraving machines allows for quick and efficient production.

As efficient and precise as laser engraving machines are, they are not without their limitations. Understanding these limitations is essential to maximize the efficiency of these machines and achieve the desired results. Thus, our further work would be aimed at enhancing the functionality of this machine and expand the range of compatibility of different materials, namely:

- Upgrading the laser that would allow working with a wider range of materials,
- Adding a rotating axis that enables the engraving of cylindrical objects,
- Adding sensors i.e. limiters (in case of a problem with the gcode and a machine moves beyond the defined range of the physical limits of its motion mechanism),
- Expanding the work area to be able to create projects with larger dimensions,
- Design and components replacement with better and more precise stepper motors, changing the construction to be lighter and easily portable.

References:

- [1] Winnie Li, Laser Engraving: The Ultimate Guide for Beginners, June 18, 2024, <https://www.xtool.com/blogs/xtool-academy/laser-engraving>.
- [2] Winnie Li, CNC Engraving vs Laser Engraving: Discover the Difference, January 04, 2023, <https://www.xtool.com/blogs/xtool-academy/cnc-engraving-vs-laser-engraving>.
- [3] Ritika Dhakate, Sakshi Gurav, Sunny Kasare and Kartikey Dhawale, A low cost mini CNC laser engraver powered by Benbox software, International Research Journal of Modernization in Engineering Technology and Science, Volume:06/Issue:05/May-2024, DOI: <https://www.doi.org/10.56726/IRJMETS56533>.
- [4] Om Kirit Vadhiyara, Vikas Sankpal, Darshan Mahesh Patel, Omprakash Ray, Sudhir Thakare, and Nehal Muchhala, CNC engraving machine, International Research Journal of Engineering and Technology (IRJET), Volume: 11 Issue: 04 | Apr 2024, e-ISSN: 2395-0056, p-ISSN: 2395-0072, <https://www.irjet.net/archives/V11/i4/IRJET-V11I458.pdf>.
- [5] Raikar Raj, Bhanushali Jay, Takarkhede Nilay, Shah Sejal and Varma Dinal, Open Source LaserGRBL – Arduino Based Laser Engraver, Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021), June 14, 2021, Available at SSRN: <https://ssrn.com/abstract=3866523> or <http://dx.doi.org/10.2139/ssrn.3866523>.
- [6] Anil Kumar J, Wilson Dsouza, Rajat Kumar, Design of CNC Machine Based Laser Engraver with Arduino UNO, Ijrasnet Journal For Research in Applied Science and Engineering Technology, ISSN: 2321-9653, 2022, <https://doi.org/10.22214/ijrasnet.2022.46249>.
- [7] Ginna Vaishnavi, D.V. Srikanth, K Nanda Kishore, B Pavan Teja, K Sandeep and C Tanusha, Design and fabrication of laser engraver, International Research Journal of Engineering and Technology (IRJET), International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 05 | May 2020, e-ISSN: 2395-0056, p-ISSN: 2395-0072, www.irjet.net.
- [8] Nguyen Van Dung, Research, design, and fabrication of laser engraving machine, International Research Journal of Modernization in Engineering Technology and Science, Volume:05/Issue:02/February-2023, e-ISSN: 2582-5208, www.irjmets.com.
- [9] Israa Fayez Yousifa, Mustafa Qahtan Alsudania, Safa Riyadh Waheeda, Zaid Nidhal Khudhaira, Myasar Mundher Adnanc, Ameer Al-khaykand, Automatic Laser Engraving Machine for Different Materials based on Microcontroller, Malaysian Journal of Fundamental and Applied Sciences, Vol. 19 (2023) 323-331, DOI: <https://doi.org/10.11113/mjfas.v19n3.2903>.
- [10] Malik Shahaan Khalid, Syed Muhammad Jaleed, Arslan Zafar, Sohaib Ali Khan, Hafiz Zia Ur Rehman, Zeashan Hameed Khan, Design and Experimental Verification of a Laser Engraving Machine, 2023 International Conference on Emerging Power Technologies (ICEPT), Topi, Pakistan, 2023, DOI: [10.1109/ICEPT58859.2023.10152428](https://doi.org/10.1109/ICEPT58859.2023.10152428).
- [11] LaserGRBL, Free Laser Engraving, 2024, <https://lasergrbl.com/>.

Quality of UML Class Diagram: An Overview

Amizic Vuk¹, Kazi Zoltan¹

¹ University of Novi Sad, Technical faculty "Mihailo Pupin", Đure Đakovića BB, Zrenjanin, Zrenjanin

vuk.amizic@tfzr.rs; zoltan.kazi@tfzr.rs

Abstract:

This review explores the evaluation of class diagram quality and the metrics used to assess the overall quality of the model. A class diagram is one of the most important diagrams created in software development. The quality measurement methods are mostly presented as factors or numbers of various states, components, class attributes, and functions. By presenting existing metrics and their definitions for evaluating the quality of class diagrams, a more comprehensive approach to creating higher-quality models. The primary source of data for scientific research papers is collected from the Google Scholar website.

Keywords:

Unified Modeling Language, class diagram, quality, evaluation

1. Introduction

The Unified Modeling Language (UML) is defined as standardized modeling language that is used in software engineering to visualize, construct and document the structure and behavior of the system that is being developed~ [1]. By providing sets of graphical notations for creating diagrams, UML modeling brings the possibility to represent different parts of software or business process. The main goal of UML is to provide tools for creation of blueprints for software systems which shows the system's design and functionality.

Each UML diagram serves the specific purpose, and class diagram's purpose is explained below:

Class diagram is used to build the structure of the system by implementing classes, attributes, class methods (functions) and relationships among classes. It is of great importance for modeling behavior of the system, and it is commonly related towards object-oriented design. [1]

2. Data and methodology

The research strategy for UML diagrams involves reviewing papers related to UML models, the quality of these models, and metrics assessing the quality of UML diagrams, all based on existing scientific literature. After data collection, the table is formed to show the amount of reviewed scientific research papers and topics they are covering. The primary source of data for the research papers was collected from the Google Scholar website. Search keywords of focus were "class", "diagram", "quality", "metrics", "uml" and "diagram". Two pages of search engine results were considered for all research papers published. Quality metrics are being evaluated for the class diagram, one of the most important diagrams that can be created in software development process.

Table 1:

Data collection based on Google Scholar

Research paper focus	Literature
Class diagram	[2], [3], [4], [5], [6], [7], [8], [9], [10]
UML	[1]
Object oriented design and metrics	[11], [12], [13]

3. Class diagram quality evaluation and metrics

The quality evaluation of class diagrams is used to show how specific UML diagrams meet various standards during their development. Evaluation is done to provide overview of how effectively and accurately they will represent the system. The goal of quality evaluation is to improve effectiveness and reliability of the UML diagrams during the software development. It is important that UML diagrams correctly represent the system's structure and behavior according to requirements as well as to cover all aspects of systems, such as classes, use cases, interactions and similar.

One of the main goals of software engineering is to assure the quality of the object-oriented software in the early phases of software life cycle [2]. By developing object-oriented software, class diagrams are used to set design and implementation of classes by improving object-oriented approaches and reduce cost revisions. By using metrics to measure class diagram quality, the overall quality of the diagram could be revised and improved. Since maintaining software drains resources, maintainability is an important factor to consider. Because many existing measures focus on code or detailed design, various measures are proposed to set structural complexity metrics for UML class diagrams, which includes the relationships such as associations, generalization, aggregation and dependencies as well as traditional metrics such as number of classes, methods and attributes [3]. Evaluating class diagrams in object-oriented software can represent the foundation for all later designs and work implementations. By having class diagram of higher quality, the overall quality of object-oriented software can be achieved [4]. By implementing object-oriented estimations, it is possible to study the judgment skills suite of a class diagram's structure and how the models have been developed and portrayed [5]. Since class diagrams helps to better understand software design, a class's importance can be measured by using specific metrics that count the number of attributes and dependencies and similar [6]. A key objective of software engineers is to continuously improve the quality of the software products and ensure robust quality assurance [7]. Due to the fact that class diagrams propose great influence on the quality of object-oriented information systems (OOIS), the quality of class diagrams is crucial to achieve overall quality of the software product [8].

Table 2:

CK Metrics [4][13][14]

Metric name	Definition
WMC	<p>The Weighted Methods per Class is defined as follows:</p> $WMC = \sum_{i=1}^n C[i]$ <p>Where c1, c2, ... cn be the complexity of the methods of a class with methods M1, M2, ...,Mn. [4]</p>
DIT	<p>The Depth of Inheritance of a class is the DIT metric for a class. In cases involving multiple inheritances, the DIT will be the maximum length from the node to the root of the tree. [4]</p>
NOC	<p>The Number of Children is the number of immediate subclasses subordinated to a class in the class hierarchy. [4]</p>

Since software architecture plays very important role through the entire software development life cycle (because it defines the system's structure), methods for measuring quality assurance are used to check the violations of design principles [9]. During the development of system architecture, it is of great importance that structural design has all necessary quality attributes which is usually done by

architecture evaluations [10]. CK Metrics, proposed by Chidamber and Kemerer are used to measure design complexity in relation to their impact on external quality attributes such as maintainability and reusability. CK Metrics asserts several metrics: the weighted methods per class, depth of inheritance of the class as well as the number of children for a class hierarchy [4].

According to [11], one of the complexity metrics for methods can be used by initially setting weight of method to 1 for each function, measuring the overall number of the functions inside of the class. According to [12], McCabe's cyclomatic complexity metric can be used to define function complexity as well.

Li and Henry's metrics are used to measure different internal attributes such as coupling, complexity and size. Those metrics were applied by Li and Henry on two real systems developed using Classic-ADA, and found that the maintenance effort could be predicted from the values of these metrics. [4]

Table 3:

Li and Henry's metrics [4][13]

Metric name	Definition
DAC	The number of attributes in a class that have another class as their type. [4]
DAC'	The number of different classes that are used as types of attributes in a class. [4]
NOM	The number of local methods. [4]

MOOD metrics combines several metrics in order to evaluate class diagram quality by measuring the use of object-oriented design mechanisms such as inheritance, measured by MIF and AIF metrics, information hiding measured by MHF and AHF metrics as well as polymorphism measured by PF metric. [4]

Table 4:

MOOD metrics [4][14]

Metric name	Definition
MHF	The Method Hiding Factor is defined as a quotient between the sum of the invisibilities of all methods defined in all of the classes and the total number of methods defined in the system under consideration. The invisibility of a method is the percentage of the total classes from which the method is not visible. [14]
AHF	The Attribute Hiding Factor is defined as a quotient between the sum of the invisibilities of all attributes defined in all of the classes and the total number of attributes defined in the system under consideration. The invisibility of an attribute is the percentage of total classes from which the attribute is not visible. [14]
MIF	The Method Inheritance Factor is defined as a quotient between the sum of inherited methods in all classes of the system under consideration and the total number of available methods (locally defined and include those inherited) for all classes. [14]

AIF	The Attribute Inheritance Factor is defined as a quotient between the sum of inherited attributes in all classes of the system under consideration and the total number of available attributes (locally defined plus inherited) for all classes. [4]
PF	The Polymorphism Factor is defined as the quotient between the actual number of different possible polymorphic situations, and the maximum number of possible distinct polymorphic situations for class C_i . [4]

Lorenz and Kidd's metrics are classified into Class size metrics, Class inheritance metrics and Class's internals metrics. Goal of these metrics is to define the static characteristics of the software design, such as usage of inheritance, amount of responsibilities in a class, etc. [4].

Table 5:
Lorenz and Kidd's metrics [4]

	Metric name	Definition
Class size metrics	PIM	This metric is used to count the total number of public instance methods in a class. Public methods are the ones that are available as services to other classes. [4]
	NIM	This metric is used to count all public, protected and private methods defined for a class' instances. [4]
	NIV	This metric is used to count total number of instance variables in a class. Instance variables include private and protected variables available to instances. [4]
	NCM	This metric is used to count total number of class methods in a class. A class method is a method that is global to its instance. [4]
	NCV	This metric is used to count the total number of class variables in a class. [4]

Table 6:

Lorenz and Kidd's class inheritance metric [4]

Class inheritance metrics	Metric name	Definition
	NMO	This metric is used to count total number of methods overridden by a subclass. A subclass is allowed to define a method of the same name as a method in one of its super-classes. This is called overriding the method. [4]
	NMI	This metric represents The Number of Methods Inherited, and it is used to count total number of any method inherited by a subclass. [4]
	NMA	This metric is used to count the total number of methods defined in a subclass. [4]
	SIX	The Specialization Index metric for each class is defined as: $\frac{(\text{NumberOfOverriddenMethods} * \text{HierarchyNestingLevel})}{\text{TotalNumberOfMethods}}$ [4]

Table 7:

Lorenz and Kidd's class internals metric [4]

Class inheritance metrics	Metric name	Definition
	APPM	The Average Parameters Per Method metrics is defined as: $\frac{(\text{TotalOfMethodsParameters})}{\text{TotalNumberOfMethods}}$ [4]

4. Conclusions

Defining quality attributes for class diagram quality measurement methods can vary and are mostly presented as factors or numbers of various states, components and general attributes such as numbers of class attributes, functions (methods, operations) and other components that are used in the inspected UML model. There is a significant lack of research papers that are covering the field of overall quality metrics and ways to measure quality attributes and evaluate the quality of other UML diagrams, like: activity diagrams, component diagrams, object diagrams, state-chart diagram and the others between those 15 diagrams that exist in UML 2.5. Most research focus towards quality evaluation and measurement metrics for class diagrams as well as use case diagrams, and that field is more covered than other UML models. As for overall metric quality of UML models, they are done individually and are intended to show specific measurement of the component that is observed, rather than giving the metric quality of a model as a whole.

References:

- [1] J. Rumbaugh, I. Jacobson, and G. Booch, "The Unified Modeling Language Reference Manual", 2021.
- [2] MM. Esperanza, M. Genero, and M. Piattini, "No-redundant metrics for UML class diagram structural complexity", Advanced Information Systems Engineering: 15th International Conference, CAiSE 2003 Klagenfurt/Velden, Austria, June 16–20, 2003 Proceedings 15, Springer Berlin Heidelberg, 2003.
- [3] M. Genero, M. Piattini, and C. Calero, "Empirical validation of class diagram metrics", Proceed. of International Symposium on Empirical Software Engineering, IEEE, 2002.
- [4] M. Genero, M. Piattini, and C. Calero, "A survey of metrics for UML class diagrams", Journal of object technology 4.9 (2005): 59-92.
- [5] B. Mathur, and M. Kaushik. "Empirical analysis of metrics using UML class diagram", International Journal of Advanced Computer Science and Applications 7.5 (2016).
- [6] F. Thung, D. Lo and MH. Osman, "Condensing class diagrams by analyzing design and network metrics using optimistic classification", Proceedings of the 22nd International Conference on Program Comprehension, 2014.
- [7] M. Genero, M. Piattini, and C. Calero, "An empirical study to validate metrics for class diagrams", Proceedings of International Database Engineering and Applications Symposium (IDEAS'02), Edmonton, Canada, 2002.
- [8] M. Genero, JA Olivas, M. Piattini and FP. Romero, "A controlled experiment for corroborating the usefulness of class diagram metrics at the early phases of OO developments", ADIS, 2001.
- [9] MRV. Chaudron, B. Katumba, and X. Ran, "Automated prioritization of metrics-based design flaws in UML class diagrams", 40th EUROMICRO conference on software engineering and advanced applications, IEEE, 2014.
- [10] S. Rajesh, and A. Chandrasekar, "An Efficient Object Oriented Design Model: By Measuring and Prioritizing the Design Metrics of UML Class Diagram with Preeminent Quality Attributes", Indian Journal of Science and Technology 9 (2016): 21.
- [11] MH. Tang, MH. Kao, and MH. Chen, "An empirical study on object-oriented metrics", Proceedings sixth international software metrics symposium, IEEE, 1999.
- [12] SR. Chidambe., and CF. Kemerer, "A metrics suite for object oriented design", IEEE Transactions on software engineering 20.6 (1994): 476-493.
- [13] BM. Goel, PK. Bhatia, "An Overview of Various Object Oriented Metrics", International Journal of Information Technology & Systems, Vol. 2; No 1: ISSN: 2277-9825.
- [14] S. Gomathi, E. Linda, "An overview of Object Oriented Metrics - A complete Survey", International Journal of Computer Science & Engineering Technology, Vol. 4; No. 09: ISSN: 2229-3345

On the Energy Consumption of Global Video Streaming

Zoran Kotevski¹, Igor Nedelkovski¹, Aleksandar Markoski¹, Andrijana Bocevska¹

¹ University "St. Kliment Ohridski" - Bitola, Faculty of Information and Communication Technologies - Bitola, Studentska bb, 7000 Bitola, North Macedonia

zoran.kotevski@uklo.edu.mk; igor.nedelkovski@uklo.edu.mk; aleksandar.markoski@uklo.edu.mk; andrijana.bocevska@uklo.edu.mk

Abstract:

In the last two decades we have witnessed how video streaming services have gained tremendous popularity, and Internet video traffic is constantly growing at a rapid pace. Certain statistics estimate that video traffic makes up to 82% of all Internet traffic, which is a quite substantial amount. If we take into account the estimation that the Internet consumed about 800 TWh of electrical energy in 2022, with projections that it will double by 2030, we come to the conclusion that video streaming services already consume more energy than countries like Brazil or Canada. It is even more surprising that in the near future it may surpass countries like Japan, Russia or even India. With the increasing demand for electrical energy, green computing and energy efficiency in video processing and transmission emerge as an important endeavor. This paper explores the current technologies, efforts and possibilities that can contribute to the reduction of energy consumption by video streaming processes. It presents a review of some of the latest research in the fields of energy efficient acquisition, coding and transmission of images and video. The paper is concluded with a review of proposed models for energy consumption in video streaming, and a proposal for future research toward quantitative estimation of video streaming energy consumption, in order to assess the potentials for mitigation of its environmental impact.

Keywords:

Video processing, video streaming, energy consumption, energy-efficient technologies

1. Introduction

Video streaming is a rapidly growing industry. Since the introduction of Macromedia Flash and the Real-Time Messaging Protocol (RTMP) in 1996 [1], a multifold of streaming services have been entering the streaming industry ever since. In the past two decades we have witnessed the huge expansion of popularity of services such as YouTube and Netflix, as well as the return to the HTTP streaming protocols like Apple's HLS [2], Adobe's HDS [3], Microsoft's Smooth Streaming [4] and MPEG-DASH [5]. In the early 2020s, mostly powered by the COVID-19 pandemics, we have experienced an enormous transformation in everyday communications and working, fueled by the explosion of video conferencing that was used in almost every aspect of human living. This video streaming trend has inspired marketers to invest greater budgets in video streaming services, while subscribers are constantly on a climb globally.

Consequently, the race of the video streaming industry to expand to new areas, and their strive to offer better Quality of Experience (QoE) of their services, led to unenviable impact on the energy consumption of global network services, which ultimately leads to negative environmental effects. This energy consumption led regulatory bodies in the European Union to impose certain regulations towards energy savings in video streaming services, which are compiled by Madlener et al. in [6].

1.1 Energy consumption estimations of video streaming

To assess the global energy consumption of video streaming services, and its environmental impact, we first present the estimations about the energy consumption of the Internet. According to Interesting Engineering [7] the estimations of the energy consumption of the Internet in 2023 ranges between 200

and 400 TWh per year. Another estimation, by Thunder Said Energy [8] claims that the Internet consumed 800 TWh of electrical energy in 2022, and the estimations for the near future are that the energy consumption of the Internet will double by 2030. Yet, one of the most worrying models is given by Jones [9], who estimates that electricity use by ICT could exceeds 20% of the global total energy consumption, and predicts an increase to 30% by 2030. If we take into account the global electrical energy consumption of 26.587 TWh in 2022, provided by Statista [10], the result would be current annual Internet energy consumption of 5.317 TWh, which is most probably overambitious. Therefore, based on the projections of Thunder Said Energy [8], for the remaining of this research we will consider an estimate of 1.000 TWh as the current annual energy consumption of the Internet.

The estimations of Internet video traffic also vary among different sources, but not as widely as previously elaborated global consumption of electrical energy. In the Cisco VNI Global IP Traffic Forecast Report [11] from 2018, the estimations are that video traffic in 2017 made about 73% of all Internet traffic, with a prediction to rise up to 82% in 2022. It is interesting to mention that this report claims that Internet video traffic grew 40% in 2017.

Similar estimations are presented by Synthesia [12] in 2024, with a claim that video made up 82% of all internet traffic in 2022, compared to 73% before the COVID-19 pandemic.

Sandvine's Global Internet Phenomena Report 2024 [13] claims that video now accounts for 70% of all Internet traffic, and in their previous report [14] they declare that data usage from video sites increases by 24% each year. It is also interesting to mention that, according to this report, Netflix, YouTube, Disney+ and Amazon Prime Video alone comprise 34% of all Internet traffic.

Another source, such as The Internet & Television Association (NCTA) [15] presents an estimation that video traffic in 2020 reached 59% of the total global network traffic.

Even though there are some variations, the aforementioned estimations clearly describe the amount of traffic that video streaming consumes and indicate the importance of tackling the issue of excessive energy consumption.

1.2 Market value estimations of global video streaming

The growth in every industry branch is driven by its economic potential, and video streaming industry is quite a representative example. According to Cloudwards report from July 2024 [16] the global video streaming industry was valued at \$50.11 billion in 2020 and rose to \$60.1 billion by 2021. The report also projects that video streaming market is expected to grow annually by 21% between 2021 and 2028, while streaming increased by 21% over the third quarter of 2020 alone.

Similar projections are given by Grand View Research report from April 2024 [17] in which the global video streaming market size is expected to reach USD 416.84 billion by 2030, registering a Compound Annual Growth Rate (CAGR) of 21.5% from 2024 to 2030. Another view of the future in video streaming is that technological advancements, like execution of Blockchain technology in video streaming and the practice of Artificial Intelligence (AI) to advance quality of videos, are expected to augment the growth of video streaming market over the forecast period.

Other estimations of the video streaming market current value, range from USD 108 billion by Statista [18], USD 670 billion by Exploding Topics [19], USD 544 billion by Forbes [20], USD 90 billion by Market Research Future for live video streaming only [21], to USD 674 billion by Fortune Business Insights [22], with growth predictions between 18% and 28% annually.

Market value estimations are in line with energy consumption estimations regarding video streaming, which supports its predicted growth in the future.

1.3 Research goal and structure

In circumstances where video streaming traffic already makes the vast majority of all Internet traffic, along with its forecasted growth, and taking into account the estimations of its energy consumption, one crucial question arises in regard to energy efficiency and sustainability. The question is: How big will the impact on energy savings be if the world switches to energy efficient video processing and streaming technologies? A simple question that is quite complex to answer. This is the main goal of our future research endeavors, but answering this question requires multitude of measures and statistics,

analyses of impact of different steps in video processing and transmission, research on novel energy-efficient technologies and modeling the energy consumption. This research represents a preliminary work, to review the current accomplishments in the field of energy efficient video streaming and to set the directions for future work. In this manner, in Section 2 we present an overview of recent advances on energy-efficient technologies concerning video processing and streaming. In Section 3 we give an overview of some of the recently proposed models for energy consumption of video streaming and we elaborate the rationale to expand these models toward realization of our research goals of quantitative estimation of energy consumption of global video streaming services. In Section 4 we conclude the paper with a reference for future work.

2. Related work on energy-efficient video streaming

More than a decade ago, Ma et al. [23] researched the technologies and possibilities to mitigate the high demands of multimedia communications on energy resources. As multimedia communications impose high resource requirements, such as high bandwidth and low transmission delay, the authors provide a state of the art, and explore energy-efficient techniques that have been proposed to date. This comprehensive survey introduced two groups of energy-efficient techniques in two integral multimedia communication phases: energy-efficient compression and energy-efficient transmission. Because of the complexity of the multimedia acquisition, compression and transmission processes, no insight into detailed energy consumption savings is given, but this comprehensive survey presents substantial information about the compression and transmission techniques of that time and provides a vision for research in energy efficiency for multimedia transmission systems.

Rajaraman et al. [24] explored the energy consumption of smartphones during video recording and streaming. Their main findings are that the majority of power is already drawn when the camera is in focus mode and not yet recording. This power is drawn by the camera internal hardware and some other hardware of the smartphone related to the video processing, and none of this hardware seems to scale the power draw with the video resolution or bit rate. The authors provide charts of power consumption in different scenarios with variety of different smartphones, and present the findings quantitatively in milli Watts (mW), which makes the results quite usable in modeling the energy consumption of video streaming systems in general. While this research provides substantial information about energy consumption in video streaming of smartphones, it does not provide insights into the energy consumption of the streaming network, nor the energy consumption of devices where the video is received, decoded and presented.

Lee et al. [25] propose an energy efficient image compression system based on discrete wavelet transform and adaptive line prediction. The proposed system includes a new bitrate control algorithm that is expected to significantly improve image quality consistency. To compare energy consumption of multimedia Internet of Things (IoT) devices with different compression algorithms, they use energy complexity method based on the number of operations in the microprocessor and the number of memory accesses. The authors claim that their proposed method achieved higher compression ratio than existing methods while maintaining lower complexity. They also claim that it demonstrated superior performance compared to existing Set Partitioning in Hierarchical Trees (SPIHT) and High Efficiency Video Coding (HEVC) method in terms of power reduction effect at system level.

Katsenou et al. [26] explore and compare the energy consumption across optimized state-of-the-art video codecs, such as SVT-AV1, VVenC/VVdeC, VP9, and x.265. From the obtained results, with the specific set of coding configurations, the authors conclude that SVT-AV1 offers the best quality-bitrate-energy tradeoff compared to the other codecs. On the other hand, for low-energy solutions they propose that x.265 is the best choice, but it comes at the cost of lower video quality.

In [27] Uitto et al. evaluate the energy and power consumption for the existing top-rated open source video encoders in proportion to video quality and bit rate. The selected video coding formats include H.264/AVC, HEVC and Google's VP9. The results presented indicate that the selected x.264 implementation of the H.264 encoder exhibits the lowest energy consumption, but the worst compression efficiency. On the other hand, x.265 implementation of HEVC appeared to have the best compression efficiency, but high energy consumption. In these experiments VP9 is the codec that provides the best results in terms of compression efficiency and energy consumption tradeoff.

Finally, Afzal et al. [28] have compiled maybe the most comprehensive survey to date, regarding the energy consumption of video streaming and its environmental impact. Some of the main research questions that this survey tries to find answers to are (i) How do various video streaming components contribute to its energy use and CO₂ emissions?, and (ii) How can we estimate or measure the energy consumption and CO₂ emissions in video streaming? They point out the most significant issues as well, among which, one of the most noticeable is the lack of a reproducible energy measurement dataset covering various device types and coding parameters.

Many research efforts concentrate on energy efficiency in image coding, mainly intended for use in energy constrained applications, such as IoT and Wireless Sensor Networks (WSN). For example, Li et al. [29] present an energy-efficient compressive image coding system, intended for use at low-energy consuming IoT. Albalawi et al. [30] proposed an energy-efficient design of the Secure Better Portable Graphics Compression (SBPG) architecture. The architecture proposed in this paper is suitable for imaging in the IoT, as the main concentration is on the energy efficiency. As one of the most important goals of current and future sensor networks is energy-efficient communication of images, Lee et al. [31] present a quantitative comparison between the energy costs associated with direct transmission of uncompressed images and sensor platform-based JPEG compression followed by transmission of the compressed image data. Furthermore, Bano et al. [32] in their research analyze image compression techniques, such as Set Partition in Hierarchical Trees (SPIHT), Set Partitioned Embedded Block Coder (SPECK), and JPEG2000 for energy constrained WSN. They also compute the energy consumption for transmitting an uncompressed 512 x 512 image from source to destination and compressed using SPIHT algorithm. Furthermore, Uthayakumar et al. [33] introduce a highly reliable and low-complexity image compression scheme using Neighborhood Correlation Sequence (NCS) algorithm. The NCS algorithm performs a bit reduction operation and then the encoding is performed to further compress the image. The proposed NCS algorithm increases the compression performance while decreasing the energy utilization of the sensor nodes with high fidelity. The authors claim that the proposed method maintains a better trade-off between compression efficiency and reconstructed image quality. Wang et al. [34], inspired by hierarchic structure of neural networks, propose a new lightweight, low-complexity learning-based image coding method entitled Green Image Codec (GIC). They claim that their experiments show that it outperforms JPEG by a significant margin and has a rate-distortion performance close to that of Better Portable Graphics (BPG) codec. Since the main compression technique in image coding algorithms like JPEG and H.265/HEVC is the Discrete Cosine Transform (DCT), Cintra et al. in [35] concentrate their research efforts to improve DCT in their proposed 8-point DCT approximations with very low arithmetic complexity. The proposed algorithms are fully compatible with modern codecs and were assessed in terms of arithmetic complexity, energy retention capability, and image compression performance. The results presented indicate very low qualitative and quantitative degradation at a considerable lower computational cost.

From the presented related research, we can conclude that the majority of research is focused on improving the energy efficiency in image coding, and a few deal with performance comparisons of existing image coding techniques. Even though energy-efficient image coding may be quite beneficial to energy constrained use cases such as in IoT and WSN, the energy savings in video streaming by sole improvement of energy efficiency in image coding alone, may not be sufficient to make an impact.

3. On the modeling of energy consumption in video streaming

In order to provide as much precise estimations on the energy consumption of video streaming systems, we need to employ mathematical modeling techniques. However, modeling the energy consumption of global video streaming is a massive task and requires detailed research on various aspects of the video streaming process, such as the coding and decoding techniques employed, the diversity of devices used in video streaming, the contribution of the different techniques used on the path of video delivery from source to destination and multitude of measures and statistics.

Several research efforts have tried to address this issue in the past. Sharab and Sarhan [36] in 2013 developed a model for power consumption in video streaming systems, focusing primarily on H.264 video encoding. For comparative purposes, they also study MJPEG and MPEG-4 video codecs. They analyze the impacts of the main H.264 video compression parameters on power consumption and

bitrate, such as quantization parameter, number of reference frames and motion estimation. One of the more important contributions of this research is that the model deals with aggregate energy consumption, taking into account the energy consumption in the capturing, encoding, and transmission phases of the live video streaming process. Yunchang Liu et al. [37] explored the energy efficiency of video streaming from a multimedia service provider perspective. This research presents a theoretical model to explore the trade-off between energy consumption and the achieved QoE for cloud based, streaming multimedia applications. The research aims to optimize the energy consumption of the cloud data center and the achieved QoE at the end users. The algorithm can guarantee desired QoE and reduce energy consumption, even without any information about the future fluctuation of user demands. Herglotz et al. [38] created a model for estimation of power consumption and QoE at the end user in a video streaming service. The research is comprehensive and provides significant insights into energy efficiency and QoE of a video streaming system, regarding the video codec, resolution, frame rate, and bitrate. The authors declare that in their measurements, possible power savings for a given device are at least 3.5%. The following Table 1 presents the key points and accomplishments of the energy consumption models concerning video streaming.

Table 1:
Comparison of models for energy consumption in video streaming

Energy consumption model	Main focus	Main results and conclusions
Sharrab and Sarhan [36]	Energy consumption in capturing, encoding and transmission of video.	Encoding for higher resolution video consumes more than 90% of the total energy consumed in the three phases. Capturing phase < 6 %. Transmission phase < 4 %.
Yunchang Liu et al. [37]	Optimization of the trade-off between the energy consumption for multimedia distribution by cloud-based data centers and QoE at end users.	The proposed optimization algorithm guarantees desired QoE and reduces energy consumption, regardless of the future fluctuation of end user demands.
Herglotz et al. [38]	Calculating the energy consumption at the client (end-user) side only.	Energy consumption at end-users depends on QoE parameters (resolution, frame rate and bitrate), and the video codec. Possible energy savings for a given device $\geq 3.5\%$.

While the aforementioned research efforts on modeling energy consumption provide substantial insights into this issue, they are focused on specific aspects of the whole video streaming paradigm. For example, the model in [36] deals only with live video streaming and takes into account the capturing, encoding, and transmission phases. The decoding phase at the end user is not taken into account. The model in [37] explores the distribution part of the whole video streaming process, i.e. the energy consumption for a cloud-based multimedia streaming application, and it omits other parts of the whole video streaming process. The model created in [38] deals with only one part of the entire streaming chain, i.e. the created model estimates the energy consumption at the end user in a video streaming service.

To answer the main question on the impact on energy savings if the world switches to energy efficient video processing and streaming technologies, we have to take the modeling paradigm a step further and encompass all the phases, technologies and devices into an energy consumption model of video streaming system. Developing such model is quite complex task. Such model will need to take into account many input elements that will require careful research prior to the creation of the model. We believe that such research will be quite valuable to convey, since energy efficiency, energy production sustainability and the reduction of greenhouse gas emissions are extremely important paradigms of today's society.

4. Conclusions

This research discusses the energy consumption of video streaming in general, and the importance of the possibilities that can contribute its reduction. It presents statistics which estimate that video streaming today probably consumes more electric energy per year than countries like Brazil or Canada, with a prospect to surpass bigger consumers like Japan, Russia or even India in the following years. A review of the latest research on technologies and possibilities to mitigate the high demands of multimedia communications on energy resources is given as well. Furthermore, a list of related models developed to estimate the energy consumption of certain stages of the video streaming process is presented, with an accent on their main focus points and their results and conclusions. The main goal in our future research is to develop a model for energy consumption of the global video streaming ecosystem, and to estimate the amount of energy savings that can be exploited if energy efficient technologies are employed. Such estimations will be quite valuable for further research on energy-efficient video streaming technologies, because they will provide invaluable information on the global energy savings, and whether future research efforts in this area are viable.

References:

- [1] Servers.com Inc. From RTMP to MRV2: The history of streaming protocols. URL: <https://www.servers.com/news/blog/from-rtmp-to-mrv2-the-history-of-streaming-protocols>. Accessed in July 2024.
- [2] HTTP Live Streaming (HLS). Apple, 2009. URL: <https://developer.apple.com/streaming/>. Accessed in July 2024.
- [3] HTTP Live Streaming (HLS). Adobe 2009 URL: https://en.wikipedia.org/wiki/Adaptive_bitrate_streaming. Accessed in July 2024.
- [4] Smooth Streaming. Microsoft Inc., 2010. URL: https://learn.microsoft.com/en-us/openspecs/windows_protocols/ms-ssstr/. Accessed in July 2024.
- [5] ISO/IEC 23009. Information technology – Dynamic adaptive streaming over HTTP (DASH), 2012. URL: <https://www.iso.org/standard/69079.html>. Accessed in July 2024.
- [6] Madlener, Reinhard, Siamak Sheykha, and Wolfgang Briglauer. The electricity-and CO2-saving potentials offered by regulation of European video-streaming services. *Energy Policy* 161 (2022): 112716.
- [7] Interesting Engineering. Internet energy usage: How the life-changing network has a hidden cost, 2023. URL: <https://interestingengineering.com/innovation/whats-the-energy-cost-internet>. Accessed in July 2024.
- [8] Thunder Said Energy. What is the energy consumption of the internet?, 2023. URL: <https://thundersaidenergy.com/2023/04/20/what-is-the-energy-consumption-of-the-internet/>. Accessed in July 2024.
- [9] Jones, Nicola. The information factories. *Nature* 561, no. 7722 (2018): 163-6.
- [10] Statista. Net electricity consumption worldwide in select years from 1980 to 2022 (2024). URL: <https://www.statista.com/statistics/280704/world-power-consumption/>. Accessed in July 2024.
- [11] Cisco Systems Inc. Cisco VNI Global IP traffic forecast 2018. URL: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_Device_Growth_Traffic_Profiles.pdf
- [12] Synthesia Ltd. 50 Video Statistics You Can't Ignore In 2024 (2024). URL: <https://www.synthesia.io/post/video-statistics>. Accessed in July 2024.

- [13] Sandvine. The Global Internet Phenomena Report 2024.
- [14] Sandvine. The Global Internet Phenomena Report 2023.
- [15] The Internet & Television Association (NCTA). New Study Examines Internet Traffic Patterns and Bandwidth Requirements (2021). URL: <https://www.ncta.com/whats-new/new-study-examines-internet-traffic-patterns-and-bandwidth-requirements>. Accessed in July 2024.
- [16] Cloudwards. 37 Streaming Services Statistics for 2024: Deep Dive Into Video & Music Streaming (2024). URL: <https://www.cloudwards.net/streaming-services-statistics/>. Accessed in July 2024.
- [17] Grand View Research. Video Streaming Market Growth & Trends (2024). URL: <https://www.grandviewresearch.com/press-release/global-video-streaming-market>.
- [18] Statista. Video Streaming (SVoD) - Worldwide (2022). URL: <https://www.statista.com/outlook/dmo/digital-media/video-on-demand/video-streaming-svod/worldwide>. Accessed in July 2024.
- [19] Exploding Topics. Video Streaming Services Stats (2024). URL: <https://explodingtopics.com/blog/video-streaming-stats>. Accessed in July 2024.
- [20] Forbes. Top Streaming Statistics In 2024 (2024). URL: <https://www.forbes.com/home-improvement/internet/streaming-stats/>. Accessed in July 2024.
- [21] Market Research Future. Live Streaming Market Overview (2020). URL: <https://www.marketresearchfuture.com/reports/live-streaming-market-10134>. Accessed in July 2024.
- [22] Fortune Business Insights. Market Research Report (2023). URL: <https://www.fortunebusinessinsights.com/video-streaming-market-103057>. Accessed in July 2024.
- [23] Ma, Tao, Michael Hempel, Dongming Peng, and Hamid Sharif. "A survey of energy-efficient compression and communication techniques for multimedia in resource constrained systems." *IEEE Communications Surveys & Tutorials* 15, no. 3 (2012): 963-972.
- [24] Rajaraman, Swaminathan Vasanth, Matti Siekkinen, and Mohammad A. Hoque. "Energy consumption anatomy of live video streaming from a smartphone." In 2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), pp. 2013-2017. IEEE, 2014.
- [25] Lee, Seong-Won, and Ho-Young Kim. "An energy-efficient low-memory image compression system for multimedia IoT products." *EURASIP Journal on Image and Video Processing* 2018, no. 1 (2018): 87.
- [26] Katsenou, Angeliki, Jingwei Mao, and Ioannis Mavromatis. "Energy-rate-quality tradeoffs of state-of-the-art video codecs." In 2022 Picture Coding Symposium (PCS), pp. 265-269. IEEE, 2022.
- [27] Uitto, Mikko. "Energy consumption evaluation of H. 264 and HEVC video encoders in high-resolution live streaming." In 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1-7. IEEE, 2016.
- [28] Afzal, Samira, Narges Mehran, Zoha Azimi Ourimi, Farzad Tashtarian, Hadi Amirpour, Radu Prodan, and Christian Timmerer. "A Survey on Energy Consumption and Environmental Impact of Video Streaming." *arXiv preprint arXiv:2401.09854* (2024).
- [29] Li, Ran, Xiaomeng Duan, Xu Li, Wei He, and Yanling Li. "An energy-efficient compressive image coding for green internet of things (IoT)." *Sensors* 18, no. 4 (2018): 1231.
- [30] Albalawi, Umar, Saraju P. Mohanty, and Elias Kougianos. "Energy-efficient design of the secure better portable graphics compression architecture for trusted image communication in the IoT." In 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 302-307. IEEE, 2016.
- [31] Lee, Dong-U., Hyungjin Kim, Mohammad Rahimi, Deborah Estrin, and John D. Villasenor. "Energy-efficient image compression for resource-constrained platforms." *IEEE Transactions on image Processing* 18, no. 9 (2009): 2100-2113.
- [32] Bano, Nishat, Monauwer Alam, and Shish Ahmad. "Energy efficient image compression techniques in WSN." In *Intelligent Communication, Control and Devices: Proceedings of ICICCD 2017*, pp. 1079-1088. Springer Singapore, 2018.
- [33] Uthayakumar, J., Mohamed Elhoseny, and K. Shankar. "Highly reliable and low-complexity image compression scheme using neighborhood correlation sequence algorithm in WSN." *IEEE Transactions on Reliability* 69, no. 4 (2020): 1398-1423.

- [34] Wang, Yifan, Zhanxuan Mei, Qingyang Zhou, Ioannis Katsavounidis, and C-C. Jay Kuo. "Green image codec: a lightweight learning-based image coding method." In *Applications of Digital Image Processing XLV*, vol. 12226, pp. 70-75. SPIE, 2022.
- [35] Cintra, Renato J., Fábio M. Bayer, Vitor A. Coutinho, Sunera Kulasekera, Arjuna Madanayake, and André Leite. "Energy-efficient 8-point DCT approximations: Theory and hardware architectures." *Circuits, Systems, and Signal Processing* 35 (2016): 4009-4029.
- [36] Sharrah, Yousef O., and Nabil J. Sarhan. "Aggregate power consumption modeling of live video streaming systems." In *Proceedings of the 4th ACM Multimedia Systems Conference*, pp. 60-71. 2013.
- [37] Liu, Yunchang, Chunlin Li, and Zhiyong Yang. "Tradeoff between energy and user experience for multimedia cloud computing." *Computers & Electrical Engineering* 47 (2015): 161-172.
- [38] Herglotz, Christian, Werner Robitza, Matthias Kränzler, Andre Kaup, and Alexander Raake. "Modeling of energy consumption and streaming video QoE using a crowdsourcing dataset." In *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1-6. IEEE, 2022.

The Social Network Facebook as a Marketing Tool in the IT Retail Sector

Mitar Jovanović¹, Olivera Stupar¹ and Mila Kavalić¹, Sanja Stanisavljev¹, Verica Gluvakov¹ Snežana Mirković²

¹ University of Novi Sad, Technical faculty "Mihajlo Pupin", Đure Đakovića bb, Zrenjanin, Serbia

² BB TRADE doo, Zrenjanin, Serbia

mitar.jovnovic@tfzr.rs; olivera.stupar@tfzr.rs; mila.kavalic@tfzr.rs; sanja.stanisavljev@tfzr.rs; verica.gluvakov@tfzr.rs; bbhrmanager@gmail.com

Abstract:

The aim of this paper is to explore how online retailers use Facebook marketing to create and maintain strong relationships with their customers. The research included an analysis of various goals set by different retailers and the strategies they employ to achieve success through the Facebook social network. The results show that retailers mainly use basic, pre-defined routines and methods that are not developed enough, focusing on commercial messages and neglecting long-term benefits, often leading to ineffective communication with consumers. Although they recognize the importance of feedback, many retailers still do not pay enough attention to analyzing and applying it, often interpreting it superficially. The paper also emphasizes the need for integrating more internet platforms, with the emergence of new and modern ones, to achieve greater success. It concludes that retailers are aware of the lack of investment in time, money, and other resources in Facebook marketing and plan to increase budgets for developing more effective strategies in the future. While planning is one thing and implementation another, we are all aware that changes happen very quickly and often when we least expect them. Those who adapt in time have a better chance of surviving in the increasingly competitive global market.

Keywords:

Social networks, Facebook, marketing, consumers, retail, IT, feedback

1. Introduction

The use of social networks and the internet has evolved to such an extent that it has become an indispensable part of everyday life. People spend an incredible amount of time connecting with others through them. The spread of social networks, especially Facebook, offers enormous opportunities for establishing more efficient and faster ways of communicating with retail consumers. Today, it is easy for consumers to access information about products and companies selling them, making them very well-informed. Facebook provides diverse opportunities for advertising and business promotion, but it is crucial to understand how to effectively use these tools in line with company goals [1]. Expressing views, personal opinions, and feelings about various products and the companies selling them to a large group of people became much easier with the advent of social network [2]. However, in this new business environment, companies find it very difficult to control complaints, criticisms, and personal opinions that users can share with others over the internet. According to Aaker (2009), this increasing power that customers now possess creates a completely new method of marketing communication through which companies actually try to establish more personal relationships with their consumers, contrary to the earlier traditional method of mass communication [3].

Urban (2003) highlighted the importance of consumer support, claiming that old and traditional marketing methods no longer fulfill their intended purpose, adding that modern consumers are now better informed and able to recognize marketing and sales tricks used by companies, which gives them better control over information that was previously the greatest advantage of companies. Urban (2003) also claims that consumers tend to find traditional marketing methods such as unsolicited letters, TV marketing (or telemarketing), and pop-up ads very annoying. He believes that these traditional marketing tools tend to increase feelings of anger or even mockery among potential consumers instead of encouraging them to buy products, which would generate profit for the company [4]. According to

Borges (2009), marketing conducted through social networks allows companies to create, maintain, and develop stronger relationships with their consumers, making mutual communication better, which further increases company profits. Based on Urban's (2003) opinions, modern retailers draw attention to how customers connect with their brands online, how they perceive them at home and in stores, while traditional retailers tend to focus on what customers think about their company's brand [5].

Among the many social networks available, Facebook stands out due to its global popularity, advertising tools and extensive user base. For IT retailers, Facebook provides not only a platform for engaging with customers but also an effective way to reach targeted audiences. The decision to focus on Facebook in this paper stems from its leading role in social media marketing and its widespread use by retailers in the IT sector in Serbia. Additionally, Facebook offers tools such as customer interaction through comments, likes, and shares, which allow for measuring engagement and marketing strategies.

2. Methodological framework of the research

2.1. Research problem

In today's digital age, interaction with consumers on social media is becoming increasingly important for building a brand and engaging the target audience [6]. In the modern business environment, companies must be agile and flexible to adapt to rapid market changes and meet growing consumer needs [7]. The main research problem of this paper is to evaluate how social networks, specifically Facebook, are utilized by IT equipment retailers to improve marketing strategies and increase sales. This includes investigating whether retailers are adequately planning and implementing their Facebook marketing campaigns, selecting trained personnel for social media management, and performing risk assessments and project planning before launching marketing initiatives. Furthermore, the paper explores how well retailers integrate consumer feedback and engagement to refine their marketing approaches.

2.2. Research subject

The main idea of this paper is to define models of how online retailers use Facebook marketing to create and maintain strong relationships with their consumers. We will consider different goals set by various online retailers regarding the use of Facebook marketing.

2.3. Research objective

According to Radosav (2023), the goal of any research is to acquire certain scientific knowledge that, in addition to social justification, also has its practicality [8]. The main objective of this paper is to examine the effectiveness of Facebook as a marketing tool from the perspective of IT retail. While the focus is on Facebook, the conclusions drawn can contribute to understanding broader social media strategies. The study seeks to identify key methods and strategies used by retailers to create, maintain, and enhance consumer relationships through Facebook, ultimately contributing to increased brand loyalty and sales.

2.4. Research hypotheses

This paper addresses one main and two additional hypotheses:

- *H: The use of the Facebook social network by IT equipment retailers increases sales compared to traditional marketing tools.*
- H1: Adequately planned/implemented Facebook campaigns increase their effectiveness.
- H2: Careful segmentation of the target audience, with the same amount of resources, achieves better effects.

2.5. Description of the research sample

This research focuses on retailers in the IT hardware industry. Retailers in the industry were selected from a list of the largest online companies in Serbia provided by "Economic Magazine" (2020). Five companies in the IT hardware industry were identified, all of which actively use Facebook pages and are primary online retailers in Serbia.

The research sample contains five of the largest online IT hardware retailers in Serbia, selected based on their active presence on Facebook and their presence in the market. These retailers were chosen to ensure that the sample is representative of the IT retail industry in Serbia, allowing for scientific conclusions to be drawn. The sample enables the study to examine how different companies with varying levels of resources and marketing sophistication utilize Facebook as a marketing platform. The selected companies are leaders in online retailing, which gives relevance and weight to the study's findings.

2.6. Place and time of research

Some of the data and information presented in this document were obtained through emails and phone calls with the staff of the mentioned companies, while most of the data were obtained from questionnaires. The research location is throughout Serbia, predominantly the Mačva region. The research time is December 2023.

2.7. Research methods, techniques and instruments

The research consists of a combination of qualitative and quantitative methods. Data collection was carried out through structured interviews with key marketing personnel from each retailer, focusing on how they use Facebook for marketing purposes. Additionally, a detailed questionnaire was designed to collect specific information on marketing strategies, engagement metrics, and consumer interaction through Facebook. The questionnaire was structured to reveal insights into the effectiveness of the campaigns and the retailers' perception of Facebook as a marketing tool. Furthermore, authors analyzed Facebook pages in order to observe the frequency and type of interactions (e.g., likes, shares, comments). This mixed method approach ensured a comprehensive understanding of Facebook's impact on marketing performance.

3. Research results and discussion

3.1. Research results

Based on the conducted research, we will analyze the results of using the Facebook social network by IT equipment retailers compared to traditional marketing methods with the aim of increasing sales.

Most respondents admitted that they want to use Facebook to promote their brand and increase brand recognition. WinWin and Tehnomanija added that they want to develop a more personal relationship with their fans and thereby increase their number. Similarly to WinWin's goal, Gigatron also hopes to use Facebook to create a more loyal relationship with consumers. GrappoloGroup plans to upgrade its Facebook profile by making it more interactive. The upgraded page will include short films, advertisements, and other user-friendly content. ComTrade sees an opportunity to use this marketing channel and take advantage of Facebook's popularity for advertising because it does not require significant investments. ComTrade explained that they avoid huge investments in Facebook marketing because the future of this social network is very unpredictable and uncertain, adding that this prevents them from setting long-term goals. However, ComTrade also added that they will have to address this issue in the future if they plan to keep pace with the competition, given that most large competing companies are already starting to transfer their consumer services and online stores.

Table 1 presents the short-term goals identified by different retailers for their Facebook campaigns. These goals indicate the strategic focus of the retailers, such as brand awareness or increasing likes.

Table 1:

Facebook campaign goals

Retailer	WinWin	Gigatron	Grappolo	Tehnomanija	ComTrade
Q1- Short-term goal	More "likes"	Brand awareness	Inform consumers	Apply a new strategy to increase "likes"	Increase sales
Q2 – Short-term goal	Increase the number of customers	Increase loyalty	Launch a larger campaign	Become a top brand in the online market	Use Facebook as a customer service channel

Regarding their presence on Facebook, all interviewed retailers responded that they had different goals when they joined online marketing through Facebook. Three out of the five interviewed retailers confirmed that they use Facebook for advertising new products and marketing campaigns. Besides Gigatron, only WinWin uses this online social network for customer service purposes, while ComTrade plans to introduce this into their online offering.

ComTrade holds the view that when a company adds consumer services and a space for exchanging opinions and discussions to their Facebook page, they must keep in mind the importance of reviewing and responding to every question or comment a consumer leaves, otherwise, the company's reputation may be jeopardized. The respondent from ComTrade added that the company's motto regarding the use of Facebook as a marketing tool is "there is no point in using social media for advertising if you are not a social company yourself." Although Tehnomanija admitted that they used Facebook only for product advertising, the respondent from this company confirmed that the company plans to publish not only interesting but also unique content on their Facebook page as part of a new marketing campaign in the near future.

The announced campaign includes special content and offers that will be available only on the company's Facebook page, which is linked to the company's official online store. In this way, the company aims to reduce the prices of their products and services and increase and transfer online traffic from their Facebook page to their official website. Three retailers confirmed that they have changed their initial goals and purposes of using Facebook, which fulfills the additional hypothesis that a well-planned/implemented Facebook campaign increases its efficiency.

Table 2:

Goals of using Facebook

Retailer	Goal and purpose of using Facebook	Does the purpose change?
WinWin	Campaign, customer services, community	No
Gigatron	Campaign, customer services, product advertising	No
Grappolo	Attract new customers, increase brand awareness	Yes
Tehnomanija	Promotion of products and services	Yes
ComTrade	Informative and marketing campaign	Yes

During the interviews, four of the respondents stated that they use two or more social media channels, while all used Facebook as a marketing tool. Only two interviewed companies have been members of Facebook for more than two years. From the standpoint of the "number of likes" most companies have around 300 "likes" except Tehnomanija, which has several thousand "likes" and ComTrade, which has nearly a thousand. Respondents offered different reasons for deciding to use Facebook. According to WinWin, any PR advertising is good advertising, while Tehnomanija followed the example of a well-known foreign company that sold similar products and achieved great success with Facebook marketing.

On the other hand, ComTrade sees Facebook as an excellent tool for informing customers, publishing company news, advertising new products, and attracting new customers. This approach in the research confirms the auxiliary hypothesis that careful segmentation of the target audience, using the same amount of resources, achieves a better effect.

Table 3 shows the number of likes and social media engagement for each retailer, reflecting their commitment to Facebook marketing and presence on other platforms.

Table 3:

Use of Social Media

Retailer	WinWin	Gigatron	Grappolo	Tehnomanija	ComTrade
Number of likes	360	250	280	3860	930
Using Facebook since	Yes, since May 2018	Yes, since June 2017	Yes, since Aug 2020	Yes, since Nov 2020	Yes, since Feb 2016
Presence on other social media	Twitter	Price comparison sites	MSN, Blocket	YouTube, Blogger	LinkedIn, Forums

The study found that while all retailers acknowledged the importance of Facebook in their marketing strategies, their levels of engagement and success varied. For instance, WinWin and Gigatron both use Facebook to increase brand awareness and build customer loyalty, but they do so with different degrees of sophistication. The larger companies, such as Tehnomanija, demonstrated more strategic use of Facebook, incorporating both promotional content and customer service functions. However, none of the retailers showed evidence of systematically using sales data to directly link Facebook activities to increased revenue. The research also highlighted that while companies recognize the potential of Facebook, many still rely on traditional one-way communication methods, limiting their ability to fully engage with consumers.

4. Research discussion

The presented research results confirm the hypothesis that the use of the Facebook social network by IT equipment retailers, compared to traditional marketing tools, increases sales. The analysis showed that the right model for applying empirical research methods was chosen, with the aim of explaining the findings in this relatively new area of marketing. The basic strategy of this paper is the combined use of inductive and deductive methods. Most companies started using Facebook because it is a cost-effective marketing tool, and the success of other companies on this platform further increased their interest.

The selected subjects primarily use Facebook pages to provide unique content, publish news, discount codes, promote innovative products, and provide consumer services [9]. To attract new consumers, Smith (2011) recommends using social media to improve consumer retention, build market share, and brand awareness [10]. It has been observed that not all respondents fully followed these recommendations and instead used Facebook more as a traditional marketing channel, rather than a channel for building relationships with consumers.

Marketing strategy allows organizations to focus their limited resources on the greatest opportunities for increasing sales [11]. Shih (2010) claims that social media and Facebook marketing should offer a

trade-oriented effect, although only one respondent provides an integrated web store on their Facebook page [12]. Most respondents have been using Facebook as a marketing tool for 4-5 years, indicating significant room for improvement. Building relationships requires time and investment [13].

ComTrade and Tehnomanija have the highest level of awareness regarding Facebook marketing and provide consumer services in line with Weber's (2009) recommendations, using Facebook as a dialogue channel. Smaller organizations have experimental approaches but lack knowledge on the proper use of Facebook. The rapidly changing business environment makes it difficult for companies to have clearly defined strategies, which may explain the absence of strategy among respondents [3].

Branding on social networks refers to dialogue and engaging users in brand development [13]. Our respondents are not yet using Facebook to its full potential for feedback, which could enhance their marketing strategies. Integrated marketing communication is important for long-term relationships [14]. Most respondents do not have a consistent approach to Facebook messaging in line with their usual marketing but promote their Facebook channel in various ways.

Porter (2009) emphasizes the importance of presence on popular platforms like Facebook. The obtained results confirm the theoretical framework, indicating that successful companies use Facebook as a marketing tool to maintain and expand market share.

It has been noted that almost all retailers express a lack of time, effort, or other resources to regularly monitor Facebook effectiveness. Periodic reviews are crucial for gathering fresh comments and feedback [15]. Although companies have employees for posting information, most do not have established routines or schedules for reviewing comments, which diminishes the potential of their Facebook strategy.

Respondents have different goals regarding Facebook marketing. Larger organizations like ComTrade and Tehnomanija have defined routines and strategies with more likes, while smaller organizations like Grappolo, WinWin, and Gigatron have fewer likes and undefined or not well-defined goals. This highlights the importance of setting high ambitions and goals for success on Facebook in terms of increasing likes and actively involving customers in brand development on the social network. Word-of-mouth advertising is more effective than traditional advertising because it is more persuasive. People are more likely to listen to their friends than an advertisement [16]. Personalization and customization of marketing messages become key success factors in the age of market segmentation and individualized consumer experience [17].

The confirmation of the hypothesis that Facebook marketing increases sales is based primarily on qualitative insights rather than strong sales data. While respondents claim that Facebook helps in boosting brand visibility, the study did not uncover direct metrics connecting Facebook usage to increased sales. This gap suggests that while Facebook is a useful tool for marketing, the full potential of its impact on sales is yet to be realized, and further studies could focus on this aspect by incorporating detailed sales analytics. Additionally, it became clear that companies often use Facebook as a traditional advertising platform, underutilizing its capacity for fostering two-way consumer relationships, which is essential for long-term marketing success.

5. Conclusion

The development of the Internet has positively influenced the communication between merchants and consumers, enabling direct interaction through web platforms. The Facebook profile is an excellent tool for collecting feedback, but respondents are reluctant to review all feedback due to the time and resources required. However, they recognize the importance of this information and plan to pay more attention to it in the future. Integrating several tools, such as newsletters and emails, has proven successful for many merchants. Facebook marketing is a useful tool for many merchants, and most recognize its positive aspects.

Larger, well-known companies set more ambitious goals and are better acquainted with the opportunities Facebook offers, focusing on two-way communication and original content. Smaller companies often set basic goals and use less developed strategies, resulting in fewer "likes" and less customer engagement in the brand identity building process. Respondents confirmed that they do not invest enough time and money in Facebook marketing but plan to increase their budgets to develop and

implement more effective strategies. Increased knowledge and awareness of the possibilities offered by Facebook marketing are changing merchants' perceptions for the better.

This study illustrates how online merchants use Facebook as a communication channel with consumers to achieve various goals. The key parameters used in the research during the given period show that the use of the Facebook social network by IT equipment retailers, compared to traditional marketing tools, increases sales, confirming the main hypothesis. There are opportunities for further research to develop more advanced strategies in this field, considering that marketing strategy allows organizations to focus their limited resources on the greatest opportunities for increasing sales.

The lack of well-designed strategies can have negative consequences for companies, so it is important to pay attention to this aspect of marketing. Given the dynamic nature of social media marketing, including Facebook, regular monitoring of strategy elements could contribute to a better understanding of their impact on the success of marketing activities on Facebook. Further research could provide new insights into communication and relationship building using this online marketing tool.

This paper has demonstrated the growing importance of Facebook as a marketing tool for IT retailers. While many companies have integrated Facebook into their marketing strategies, there is significant room for improvement, particularly in developing more interactive and consumer-focused campaigns. The research confirms that the effective use of Facebook can increase brand recognition and engagement, although the direct impact on sales remains unclear. The main contribution of this paper lies in its identification of gaps in current practices and the suggestion that retailers must move beyond one-way communication in order to recognize the benefits of social media marketing.

References:

- [1] D. A. Aaker, Strategic market management, 9th ed., Wiley, 2009.
- [2] D. A. Aaker, Building Strong Brands, Simon & Schuster, 2019.
- [3] J. Berger, Contagious: How to Build Word of Mouth in the Digital Age, Simon & Schuster, 2016.
- [4] B. Borges, Marketing 2.0: Bridging the Gap between Seller and Buyer through Social Media Marketing, 1st edition, Tucson, USA 2009.
- [5] O. Holm, Integrated marketing communication: from tactics to strategy, Corporate Communications, An International Journal, Vol. 11 Iss: 1, pp.23 – 33, 2006.
- [6] F. Kotler, G. Armstrong, Principles of Marketing, Pearson, 2021.
- [7] F. Kotler, K.L. Keller, Marketing Management, Pearson, 2022.
- [8] M. Porter, Strategy and the Internet, Harvard Business Review, 2009.
- [9] D. Radosav, Metodologija naučno-istraživačkog rada, Zrenjanin: Technical Faculty “Mihajlo Pupin”, 2023.
- [10] C. Shih, The Facebook Era: Tapping online Social Networks to Market, Sell and Innovate, Pearson Education, 2010.
- [11] M. Smith, The New Relationship Marketing: How to Build a Large, Loyal, Profitable Network Using the Social Web, John Wiley & Sons, 2011.
- [12] G.L. Urban, Digital Marketing Strategy: Text and Cases, Prentice Hall, 2003.
- [13] A. Vahl, J. Haydon, J. Zimmerman, Facebook Marketing All-in-One For Dummies, Wiley, 2013.
- [14] L. Weber, Marketing to the Social Web: How digital customer communities build your business, Wiley & Sons, 2009.
- [15] Ekonomski magazin (2020). Retrieved January 2024, from <https://alsat.mk/mk/category/tv-alsat-2/emisii/ekonomski-magazin/>
- [16] <http://www.comtradeshop.com/>
- [17] <http://www.gigatronshop.com/>
- [18] <http://www.grappologroup.com/>
- [19] <http://www.winwin.rs/>
- [20] <http://www.tehnomanija.rs/>

The impact of Artificial Intelligence on personalization strategies in digital marketing

Verica Gluvakov¹, Mila Kavalic¹, Mihalj Bakator¹, Stefan Ugrinov¹, Sanja Stanisavljev¹ and Snežana Mirković²

¹ University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

² BB TRADE doo, Zrenjanin, Serbia

verica.gluvakov@tfzr.rs; mila.kavalic@tfzr.rs; mihalj.bakator@tfzr.rs; stefan.ugrinov@tfzr.rs; sanja.stanisavljev@tfzr.rs; bbhrmanager@gmail.com;

Abstract:

This paper explores the impact of artificial intelligence (AI) on personalization strategies in digital marketing, with a focus on AI's ability to enhance customer engagement, improve marketing efficiency, and drive conversions. AI enables businesses to analyze large datasets in real time, allowing precise targeting and dynamic customer interactions. The paper identifies key advantages, including increased conversion rates and brand loyalty, while acknowledging challenges related to data privacy and ethical considerations. The future of AI-driven marketing will be shaped by technologies like augmented reality and voice tools. The paper concludes that while AI offers significant potential in digital marketing, a balanced approach that respects consumer privacy and autonomy is essential for sustainable growth.

Keywords:

AI, digital marketing, personalization, marketing efficiency, customer engagement

1. Introduction

Artificial intelligence (AI) has become a transformative force in digital marketing, particularly in the area of personalization strategies. The integration of AI into marketing practices not only streamlines processes but also enhances the precision with which businesses can connect with individual consumers [1]. One of the most significant impacts AI has made is its ability to automate complex marketing tasks. AI technologies such as machine learning, natural language processing, and predictive analytics are capable to handle large volumes of data, allowing marketers to make data-driven decisions with greater speed and accuracy. As a result, brands are able to create more dynamic marketing strategies, ensuring that they can stay competitive in the marketplace [2]. Personalization strategies powered by AI can be seen in a variety of applications, from tailored email campaigns and customized product recommendations to dynamic ad targeting. Consumers are increasingly expecting these personalized interactions and businesses that effectively use AI to meet these expectations often experience improved engagement, higher conversion rates, and stronger customer loyalty [3].

The primary motivation of this paper is to explore AI's contribution to creating customer-centric marketing strategies. The methods analyzed include AI-driven tools like recommendation engines and dynamic ad targeting. This paper contributes to the literature by highlighting the potential of AI to refine personalization strategies and examines the associated risks of privacy and ethical concerns. Second section of the paper discusses the AI tools used in marketing, third section outlines challenges and ethical implications, and fourth section proposes a model for improving AI-personalized marketing.

2. AI-driven personalization in digital marketing

The application of AI in digital marketing includes a variety of tools and technologies that assist marketers in analyzing large amounts of data, predicting consumer behavior, and automating routine tasks. AI systems are capable of processing large datasets more quickly and accurately than traditional methods, allowing businesses to draw insights that would otherwise remain hidden. This allows

businesses to deliver targeted ads, personalized recommendations, and tailored content to individual consumers. Machine learning, a subset of AI, is particularly useful in this regard. Through machine learning, algorithms identify patterns in customer data and forecast future behavior, enabling marketers to anticipate customer needs and make data-driven decisions. Natural language processing (NLP), another AI technology, enhances marketing by enabling tools like chatbots and sentiment analysis. Chatbots can automate customer service interactions, providing personalized responses based on prior interactions, improving user experience and efficiency [4].

As AI becomes increasingly integrated into digital marketing, its ability to automate tasks has been a major driver of its adoption. AI tools are now frequently used to automate customer segmentation, email marketing, ad targeting, and content recommendation. These tasks, which would otherwise require significant manual effort and time, can now be performed with precision and at scale through AI systems. AI-powered tools can automatically optimize digital ad placements by analyzing performance data in real-time, adjusting bids and targeting settings to maximize return on investment (ROI). The ability to automate these complex processes not only saves time but also ensures that marketing efforts are consistently refined to achieve better outcomes [5].

In today's digital landscape, consumers expect brands to understand their preferences and deliver relevant content that meets their needs. AI enables this level of personalization by analyzing a wide range of data points, such as past purchases, search queries, and online behavior to build detailed customer profiles. These profiles allow marketers to predict individual consumer preferences and deliver highly customized recommendations, whether it be product suggestions, personalized emails, or targeted ads. This approach helps brands create more meaningful connections with their customers, resulting in higher engagement and increased loyalty [6].

Personalization through AI can be seen across various touchpoints in digital marketing. For example, AI is commonly used in recommendation engines, which analyze user data to suggest products or services based on previous interactions. This not only enhances the customer experience but also increases the likelihood of repeat purchases or continued engagement. Similarly, AI is used in personalized email campaigns, where the content and timing of messages are tailored to individual recipients. AI can analyze factors such as open rates, click-through rates, and purchase history to determine the optimal time to send an email and what content is most likely to resonate with the recipient [7].

AI's role in ad targeting is another key aspect of personalization. Through predictive analytics, AI can determine which ads are most relevant to specific users based on their online behavior and interactions with past campaigns. Predictive models analyze historical data to predict future customer actions, such as purchasing trends or content preferences. This helps marketers create personalized campaigns that resonate with specific customer segments [8].

3. Challenges and future aspects of AI-personalization in digital marketing

Despite the potential of AI, several challenges must be addressed, primarily in data privacy and ethical use. Personalization relies on the collection and analysis of large amounts of personal data, including consumer preferences, browsing behaviors, purchase history, and social media interactions. This data is essential for AI systems to generate accurate predictions and create tailored marketing strategies. However, the increasing use of such personal information raises concerns regarding data security and how companies handle this sensitive information. Many consumers are careful in sharing their data, particularly in light of high-profile data breaches and scandals involving the misuse of personal information [9].

A key challenge that businesses face when using AI for personalization is ensuring compliance with data protection regulations. Laws such as the General Data Protection Regulation (GDPR) in the European Union set strict guidelines for how personal data should be collected, processed, and stored. Compliance with data protection regulations, such as the General Data Protection Regulation (GDPR), is essential. Companies must ensure transparency in data collection and allow users to opt out of AI-driven personalization if desired. Misuse of data can result in legal penalties and damage brand trust. A perceived lack of transparency in data practices can lead to a loss of trust, which can, in turn, impact brand loyalty and customer retention [10].

In addition, while personalization is designed to provide more relevant and helpful experiences for consumers, it can sometimes cross a line into manipulation. AI systems have the ability to predict consumer behavior with increasing accuracy, often anticipating needs before the consumer is even aware of them. While this can improve customer satisfaction in many cases, it also raises ethical questions about how much influence companies should have over consumer choices [11].

However, despite these challenges, the future of personalization in digital marketing with AI holds large potential, driven by continuous advancements in technology. One of the most promising developments in this field is the rise of hyper-personalization which relies on AI to analyze behavioral data in real time, allowing brands to deliver highly customized experiences instantly. This level of personalization not only increases the likelihood of conversion but also strengthens the relationship between the brand and the consumer, as it demonstrates a deep understanding of individual needs and preferences [12]. The integration of AI with other emerging technologies such as augmented reality (AR) and virtual reality (VR) also opens up new possibilities for personalized marketing. These immersive technologies enable brands to create highly interactive and engaging experiences that are tailored to individual users [13].

AI-driven tools are becoming increasingly sophisticated, allowing for more natural and personalized interactions with customers. Chatbots, for example, can engage in real-time conversations with users, offering personalized product recommendations, answering queries, and guiding users through the purchasing process. As these tools become more advanced, they can learn from past interactions to provide even more relevant and helpful responses, making the customer experience smoother and more enjoyable [14]. Virtual assistants like Amazon's Alexa and Google Assistant are also playing a growing role in personalized marketing, as they can use AI to recommend products, services, or content based on the user's preferences and voice search history. As AI becomes better at understanding and interpreting human behavior, marketing strategies will likely become even more personalized, offering unique experiences tailored to the specific wants and needs of each individual consumer.

4. Theoretical model for improving AI personalization in digital marketing

To develop a theoretical model for improving AI personalization in digital marketing, three main elements can be identified. These elements represent the core components of the AI-personalization interaction in marketing. Each element is supported by several sub-elements that describe specific aspects of how AI influences personalization strategies. These elements and their sub-elements are presented visually in Figure 1. The model illustrates the flow from data collection through AI-driven personalization efforts to improved consumer interaction.

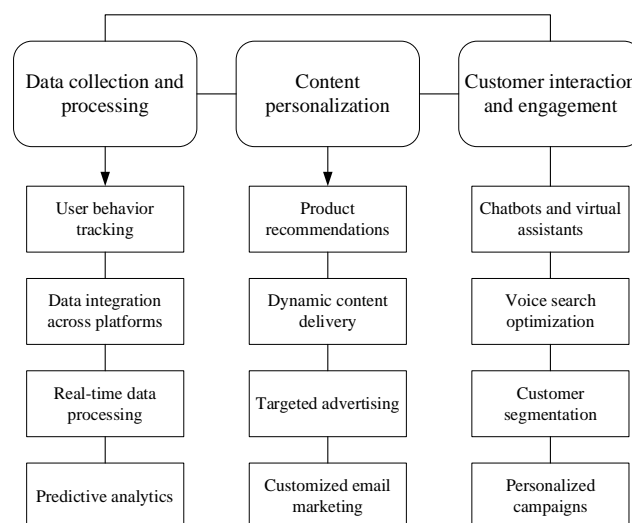


Figure 1: Theoretical model for improving AI personalization in digital marketing

The relationship between the sub-elements and the main elements in presented model is highly interconnected. Each sub-element plays a specific role within its main element, contributing to the

overall function of AI-driven personalization. The interaction between data collection and processing, content personalization, and customer interaction and engagement forms a cohesive system, where each sub-element influences the other to create a more personalized and effective marketing strategy.

In the first main element, data collection and processing, the sub-elements such as user behavior tracking and data integration across platforms work together to provide a detailed picture of the consumer. User behavior tracking allows AI systems to capture key insights into what the consumer is doing online—whether browsing, clicking on specific links, or spending time on particular pages. This data is invaluable because it serves as the raw material for personalization. The more granular the data, the better AI can understand individual preferences. Data integration across platforms enhances this by combining information from various sources, such as social media interactions, search histories, and in-app behaviors. This unification ensures that the consumer profile is complete, preventing gaps in data that could hinder accurate personalization. Real-time data processing further strengthens the element by allowing AI to analyze and act on information as it is collected, leading to immediate adjustments in marketing strategies. Predictive analytics ties everything together by using past data to forecast future actions, enabling marketers to anticipate consumer needs and deliver personalized recommendations before the customer even expresses interest.

The second main element, content personalization, relies heavily on the data provided by the first element. The relationship between data processing and content personalization is foundational to the success of AI-driven marketing strategies. Product recommendations, for instance, are directly influenced by the insights gained from user behavior tracking and predictive analytics. When AI identifies patterns in consumer behavior, it can suggest products that align with individual preferences, leading to higher chances of engagement and purchase. Dynamic content delivery, another sub-element, benefits from real-time data processing, as AI can alter the content displayed on a website or ad in response to real-time user behavior. For example, if a user shows interest in a specific product, AI can immediately adjust the content to promote similar items, ensuring a relevant and engaging experience. Targeted advertising is similarly influenced by the data collection processes. AI uses information about consumer segments and preferences to deliver ads that are more likely to resonate with specific users. This prevents irrelevant ads from being shown, reducing ad fatigue and increasing the effectiveness of marketing campaigns. Customized email marketing also builds on the insights provided through data integration and segmentation, ensuring that each email sent is uniquely tailored to the recipient's interests and behaviors, increasing the likelihood of conversion.

The third main element, customer interaction and engagement, is where the influence of AI becomes most visible to the consumer. The success of AI-driven chatbots and virtual assistants, for example, is contingent upon accurate data collection and processing. Chatbots rely on historical customer data to provide personalized responses, helping to create a seamless interaction. They can answer questions based on previous purchases or guide the customer through the buying process by using predictive analytics to recommend the next steps. Voice search optimization, another key sub-element, is closely tied to user behavior tracking and data integration. When a user performs a voice search, AI draws on past interactions and behaviors to deliver results that are more personalized and relevant. This improves the user experience, making the customer feel understood and valued. Customer segmentation, which is facilitated by AI's ability to analyze vast amounts of data, plays a significant role in determining how brands interact with different segments of their audience. Accurate segmentation ensures that the right message is delivered to the right audience, enhancing engagement by targeting consumers with content that speaks directly to their needs and preferences. Finally, hyper-personalized campaigns bring all these sub-elements together, leveraging real-time data processing, predictive analytics, and dynamic content delivery to create marketing messages that are uniquely tailored to each individual at the exact moment they are most likely to engage.

The flow of influence between these sub-elements and main elements is circular and continuous. Data collection and processing inform content personalization, which in turn enhances customer interaction and engagement. Customer responses to these personalized interactions feed back into the data collection process, allowing AI to refine and improve future personalization strategies. Each component relies on the other, forming a comprehensive system that leverages AI's ability to analyze, predict, and act on consumer data, ultimately resulting in more efficient and effective digital marketing strategies.

5. Suggestions and guidelines

Based on analyzed theory and presented model, the following guidelines and recommendations are provided:

- Implement ethical guidelines for AI usage in marketing - Governments and industry bodies should establish clear ethical frameworks to guide AI use, ensuring transparency, fairness, and accountability in marketing practices. These guidelines should address issues like consumer data privacy, consent, and minimizing biases in AI-driven personalization efforts.
- Promote consumer transparency and trust through data privacy policies - Businesses should make data collection practices transparent by informing consumers how their data will be used in AI-driven personalization. Offering clear opt-in and opt-out options can build trust, enhance brand reputation, and improve customer relationships.
- Adopt real-time data processing for hyper-personalized marketing - Enterprises should invest in AI-driven tools that analyze consumer behavior in real-time, allowing for more dynamic and contextually relevant marketing campaigns. This approach can improve engagement and conversion rates, as consumers receive personalized offers at the right time and context.
- Use predictive analytics to anticipate market trends and customer needs - Enterprises should harness AI's predictive capabilities to analyze data trends and forecast future consumer preferences and market shifts. This allows businesses to stay ahead of competitors, innovate more effectively, and deliver more relevant marketing strategies tailored to evolving demands.
- Develop AI-driven customer service tools for better engagement - Enterprises should implement AI-powered chatbots, virtual assistants, and personalized messaging systems to enhance customer service and improve engagement. These tools can offer real-time assistance and customized solutions, leading to more positive customer experiences.

6. Conclusion

The integration of artificial intelligence into digital marketing has significantly transformed how businesses approach personalization strategies. AI's ability to collect, process, and analyze vast amounts of consumer data in real-time allows marketers to deliver highly tailored content and experiences that resonate with individual users. This capability has enhanced customer engagement, improved conversion rates, and strengthened brand loyalty by creating more relevant and personalized interactions. The rise of AI in marketing has also introduced efficiencies in processes such as customer segmentation, ad targeting, and content recommendations, making marketing campaigns more precise and dynamic. However, the increasing reliance on AI for personalization presents challenges related to data privacy, ethical considerations, and algorithmic biases. As AI systems become more sophisticated, concerns over how personal data is collected, used, and protected have grown, prompting the need for stricter regulatory frameworks and transparency from businesses. Ethical concerns, such as the potential for manipulation in hyper-personalized marketing and biases in AI algorithms, require careful attention from enterprises to ensure that AI is used responsibly. Maintaining a balance between leveraging AI's capabilities and respecting consumer privacy and autonomy is crucial for the sustainable growth of AI-driven marketing strategies. Looking to the future, AI's role in digital marketing will continue to evolve as emerging technologies such as augmented reality, virtual reality, and voice-activated tools are integrated into personalized marketing campaigns. These innovations will provide even more immersive and interactive experiences for consumers, enhancing engagement and further refining personalization efforts. At the same time, governments, enterprises, and individuals must collaborate to develop ethical guidelines, invest in education and skills development, and ensure that AI is used in a way that benefits both businesses and consumers. The continued advancement of AI in marketing holds immense potential, but it also requires a commitment to ethical practices and responsible data management to fully realize its benefits.

References:

- [1] Van Esch, Patrick, and J. Stewart Black, Artificial intelligence (AI): revolutionizing digital marketing, *Australasian Marketing Journal*, 29.3, 2021, pp. 199-203.
- [2] M. Sinha, J. Healey, T. Sengupta, Designing with AI for digital marketing, Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization, 2020.
- [3] Z. Christos, M. Vlachopoulou, Artificial intelligence in digital marketing: Insights from a comprehensive review, *Information*, 14.12, 2023, pp. 664.
- [4] Arun Kumar, B. R., AI-based digital marketing strategies—a review, *Inventive Computation and Information Technologies: Proceedings of ICICIT 2020, 2021*, pp. 957-969.
- [5] N. Kiran, R. Gupta, Application of AI technology in modern digital marketing environment, *World Journal of Entrepreneurship, Management and Sustainable Development*, 17.3, 2021, pp. 318-328.
- [6] A. Daher Raddad et al., Exploring the impact of artificial intelligence in personalized content marketing: a contemporary digital marketing, *Migration Letters*, 20.S8, 2023, pp. 548-560.
- [7] B. Sodiq Odetunde et al., The role of AI in marketing personalization: A theoretical exploration of consumer engagement strategies, *International Journal of Management & Entrepreneurship Research*, 6.3, 2024, pp. 936-949.
- [8] A. Farseev, Under the hood of social media advertising: how do we use AI responsibly for advertising targeting and creative evaluation, *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023.
- [9] D. Kumar, S. Nidhi, Ethical and legal challenges of AI in marketing: an exploration of solutions, *Journal of Information, Communication and Ethics in Society*, 22.1, 2024, pp. 124-144.
- [10] J. E. Bessen et al., *GDPR and the Importance of Data to AI Startups*, NYU Stern School of Business, 2020.
- [11] Y. Su, E. Wang, P. Berthon, Ethical marketing AI? A structured literature review of the ethical challenges posed by artificial intelligence in the domains of marketing and consumer behavior, *Proceedings of the 56th Hawaii international conference on system sciences 2023*, 2023.
- [12] B. Singh, C. Kaunert, *Future of Digital Marketing: Hyper-Personalized Customer Dynamic Experience with AI-Based Predictive Models, Revolutionizing the AI-Digital Landscape*, Productivity Press, 2024, pp. 189-203.
- [13] A. Sharma, A. Bansal, Digital marketing in the metaverse: Beginning of a new era in product promotion, *Applications of Neuromarketing in the Metaverse*, IGI Global, 2023, pp. 163-175.
- [14] A. Nirwana, Implementation of artificial intelligence in digital marketing development: A thematic review and practical exploration, *Jurnal Manajemen Bisnis, Akuntansi Dan Keuangan*, 2.1, 2023, pp. 85-112.

Algorithmic Management in the Platform Economy

Ivana Simić¹ and Lucija Stojković²

¹ Faculty of Economics, University of Niš, Trg kralja Aleksandra Ujedinitelja 11, Niš, Republic of Serbia

² Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 4, Niš, Republic of Serbia

ivana.simic@eknfak.ni.ac.rs; lucija.stojkoviccc@elfak.rs

Abstract:

This paper explores the concept of algorithmic management within the context of the platform economy, focusing particularly on the roles that algorithmic management plays in platform-based businesses. By examining how algorithms manage tasks, monitor worker performance, and regulate behavior, the paper provides a comprehensive analysis of the complexity of managerial functions traditionally performed by humans. Ultimately, the analysis aims to clarify the implications of algorithmic systems for both platforms and workers, contributing to a deeper understanding of their influence on labor practices and economic organization.

Keywords:

Algorithmic management, platform economy, algorithms, digital technology.

1. Introduction

The intensive use of contemporary digital technologies in various economic activities, along with operations based on vast amounts of data, has led to the relatively new concept of algorithmic management. Although this concept predominantly focuses on the algorithms themselves and the data that feed them, rather than on the medium or space used for coordination [1], a comprehensive analysis inevitably requires consideration of this secondary aspect as well. In this regard, Sara Baiocco and her colleagues particularly emphasize the importance of examining the concept of algorithmic management “in the context of the impact of the digital revolution on work and employment” [1].

Namely, based on Eurofound’s [2] argument that the digital revolution affects work and employment through three vectors of change - automation, digitization, and platformisation - Baiocco and her colleagues observe that although algorithmic management is linked to all three vectors, it is particularly associated with the last one, platformisation [1]. They explain this by noting that the concept of platformisation “is the closest to algorithmic management,” stating that “in all digital platforms... there will be some algorithmic management embedded” and that “in most instances of algorithmic management, there will be some type of digital platform” [1].

Accordingly, the aim of this paper is to contribute to clarifying the nature of the relationship between digital platforms and algorithmic management, primarily through an analysis of the roles of algorithmic management within these platforms. To achieve this aim, the paper is structured with an Introduction and Conclusion, as well as sections analyzing the context of the platform economy (The Platform Economy Context), the concept of algorithmic management (The Concept of Algorithmic Management), and the main roles that algorithmic management performs within the platform economy (The Roles of Algorithmic Management).

2. The Platform Economy Context

The increasing development of digital-driven technologies (Internet, cloud computing, big data, Internet of Things, etc.) over the last few decades has accelerated the emergence of a new form of economic system supported by digital platforms and online marketplaces, known as the platform economy [3].

Eurofound refers to platforms as “new forms of economic organization” [2] and defines them as “digital networks that coordinate transactions in an algorithmic way” [2]. Two key elements of this definition, and thus of platforms, can be identified. Firstly, the networks, as structured digital ‘spaces’ where goods or services can be offered or requested, systematically collect, organize, and store large amounts of data about platform users and their transactions. Secondly, a set of algorithms within the platforms align and coordinate transactions in an automated way [2].

Similarly, Sara Baiocco and her colleagues describe platforms as providing “digital environments for all kinds of interactions,” representing “spaces where vast amounts of data can be collected from several sources (websites, internet-based devices such as mobile phones, and so on), as well as from users such as workers, clients, and customers using trackers and other digital tools” [1].

Analyzing the key characteristics of platforms, Peter C. Evans and Annabelle Gawer particularly emphasize two closely related aspects. One highlights the presence of network effects within platforms, while the other underscores their digital nature [4]. Regarding the first aspect, Evans and Gawer assert that “network effects are prevalent in platforms” and that platforms become more valuable and attractive to potential new users as more people use them. On the other hand, the digital nature of platforms signifies that platforms “capture, transmit, and monetize data, including personal data, over the Internet... and have a software engine at their heart” [4].

David Stark and Ivana Pais point out that the nature of relationships involving platforms is three-sided, encompassing platform owners, providers (those who offer services or products), and users/consumers (those who utilize these services or products) [5]. Within this relationship, one of the key challenges is managing the various activities on the platform. Stark and Pais see the solution to this challenge in algorithmic management [5].

3. The Concept of Algorithmic Management

Algorithmic management, also known as “management-by-algorithm” [6], has emerged as a defining characteristic of the platform economy [7]. Min Kyung Lee and his colleagues [8] were among the first to define algorithmic management, describing it as “software algorithms that assume managerial functions and the institutional devices that support these algorithms in practice” [8]. Alex J. Wood defines it as “the use of software algorithms to automate organizational functions traditionally carried out by human managers” [9]. Mohammad Hossein Jarrahi and his colleagues provide a more concise definition, describing algorithmic management as “the delegation of managerial functions to algorithms” [6]. Essentially, algorithmic management is a management approach primarily associated with platform-based businesses [9, 6].

In the context of the platform economy, Mareike Möhlmann and her colleagues define algorithmic management as “the large-scale collection and use of data on a platform to develop and improve learning algorithms that carry out coordination and control functions traditionally performed by managers” [10]. Möhlmann et al. highlight two key dimensions of algorithmic management in online platforms: matching and control [10]. For instance, Alexander Benlian and his colleagues illustrate these dimensions using the digital platform Uber. Uber employs AI-based algorithms to match drivers with customers and to control (i.e., direct, evaluate, and reward/sanction) drivers’ work behaviors [7].

Matching, a dimension of algorithmic management that facilitates the alignment of suppliers and consumers through algorithms or the synchronization of supply and demand on a digital platform, is also highlighted by Sara Baiocco and her colleagues [1]. The second dimension suggested by Baiocco et al. is coordination, which involves managing all activities to achieve the platform’s objectives [1]. According to Baiocco et al., through matching and coordination, algorithmic management performs functions typically handled by human managers in traditional organizations, such as planning, organizing, leading, and controlling [compare with: 1]. It is important to note that, in digital platforms, planning remains at least partially assisted by algorithms, as key strategic business decisions are still largely made by human managers [1].

4. The Roles of Algorithmic Management

It is evident that algorithmic management systems have become increasingly complex, incorporating various functions to oversee and direct workers as well as activities within organization. Alexandra Mateescu and Aiha Nguyen identify five key aspects of algorithmic management, which are summarized as follows [11]:

- Prolific data collection and surveillance (technology enables extensive monitoring of workers);
- Real-time responsiveness (data informs immediate management decisions);
- Automated or semi-automated decision-making (decisions are often made with minimal human intervention);
- Performance evaluations (metrics and rating systems are used to assess worker performance);
- Behavioral incentives (“nudges” and penalties are employed to influence worker behavior).

Analyzing the use of algorithmic management practices across various work environments, Alex J. Wood particularly emphasizes its role in [9]:

- Algorithmic direction (determining tasks, order, and timing);
- Algorithmic evaluation (using reputational systems to rank workers);
- Algorithmic discipline (implementing mechanisms of punishment and reward to ensure compliance).

Algorithmic direction of workers. Algorithmic management involves the automatic allocation of tasks to workers through handheld devices, smartphones, and computers. For example, Wood highlights ride-hail drivers in the U.S. (Uber, Lyft) who, once they activate their smartphone app, receive trip requests with a 15-second window to accept. If they accept, the app provides the passenger’s location via a map display. The limited time given by the algorithm to reject a request effectively forces drivers to accept the assigned work. Additionally, the app (or algorithm) directs the drivers on the most efficient route to reach the customer’s destination [9].

Algorithmic evaluation of workers. Using the example of ride-hail drivers in the U.S., whose practices have been studied by certain researchers [8, 12], Wood explains this function of algorithmic management. According to Wood, passengers rate drivers on a scale of one to five stars, and drivers receive weekly performance metrics via an app on their smartphone. These apps also collect personalized data on braking and acceleration speeds to further assess driving performance. Additionally, the algorithm can recommend that drivers exhibiting erratic behavior take a break [9].

Algorithmic discipline of workers. Studies [8, 12] on the ride-hailing platforms Uber and Lyft in the U.S., as cited by Wood [9], reveal that drivers with low customer ratings are automatically deactivated from the platforms, preventing them from continuing to work. If a driver receives a poor rating or is deactivated due to low ratings, they have the option to contact Uber to request a review of the rating or the deactivation decision. In some cities, drivers can also choose to complete a “quality improvement course” at their own expense to regain access to the platform [9].

By concentrating on the practice of algorithmic management within selected digital labor platforms (including online freelancing, microtasking, competitive programming, and location-based taxi and delivery services), Sara Baiocco and her colleagues highlight the following key functions of algorithmic management [1]:

- Algorithmic planning and strategizing;
- Algorithmic allocation of work and direction;
- Algorithmic control and monitoring of workers.

Although planning and strategizing functions on digital platforms are predominantly managed by humans, algorithms are at least partially employed on certain ride-hailing and delivery platforms. In these cases, the real-time data generated by drivers provides a significant competitive advantage, enabling platforms to optimize their pricing strategies. By analyzing fluctuations in demand and supply during peak hours, as well as traffic patterns, these platforms can utilize algorithms to adjust prices instantaneously, a practice commonly referred to as “surge pricing” [1].

The situation differs significantly for managerial tasks such as work allocation and direction, which are primarily executed without human intervention, primarily due to the data collected by specific platforms. For example, taxi platforms utilize this data to match customers requesting rides at particular locations with drivers within minutes. The algorithm assigns tasks to drivers based on factors such as estimated time of arrival, customer ratings, and the driver's history of ride cancellations or acceptances. Drivers typically have 15 to 40 seconds to decide whether to accept or decline a ride, based on limited information about the passenger's destination, which complicates the assessment of the ride's profitability. Additionally, platforms use tracked data to determine and suggest optimal routes for drivers, aiming to minimize service time by considering real-time traffic conditions via GPS data [1].

Digital labor platforms also facilitate continuous control and monitoring of workers [13]. Baiocco and her colleagues illustrate this aspect of algorithmic management using an example from freelance platforms. They explain that these platforms offer clients a range of tools for communication, management, and ongoing monitoring of work progress. Workers are often required to meet specific software and hardware requirements outlined in the platform's terms of service. For instance, Upwork provides hourly contract workers with a "work diary" feature that, when activated, records the number of hours worked, keystrokes made, and captures random screenshots (six times per hour) during project tasks. This data is accessible to clients, allowing them to monitor worker activity, track working hours, and oversee progress in real time [1].

5. Conclusions

Algorithmic management has fundamentally reshaped the platform economy, driving changes in how work is organized and managed. While these systems offer platforms efficiency through automated decision-making and real-time monitoring, they also raise critical concerns around fairness, worker autonomy, and transparency. Platforms, by relying on algorithms to direct and evaluate workers, have created a system where human oversight is often minimized, leading to new challenges in ensuring accountability. As algorithmic management continues to evolve, it becomes increasingly important to address the ethical implications and to consider ways of incorporating human judgment alongside algorithmic processes to create more balanced, equitable systems for workers.

Some limitations of this paper include its focus on a selected group of platform-based industries, which may limit the generalizability of the findings to other sectors. Additionally, the rapidly changing nature of digital platforms and the continuous evolution of algorithmic technologies may make some observations temporary or context-specific.

Despite these limitations, the paper offers several benefits for academic researchers and experts in the economy. It provides a comprehensive framework for understanding the mechanisms of algorithmic management and its impact on labor practices, contributing to a deeper analysis of digital economies and managerial automation. For practitioners, it highlights the operational advantages and challenges posed by algorithms, which can influence future strategies for managing platform-based workforces.

Possible further research directions could explore the long-term effects of algorithmic management on worker well-being, particularly in relation to psychological and social factors. Additionally, comparative studies across different platforms and industries would provide valuable insights into how algorithmic management functions in various contexts as well as its broader societal implications.

Acknowledgment:

The paper is a result of research based on the Agreement (number 451-03-65/2024-03) signed in 2024 by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia and the Faculty of Economics, University of Niš.

References:

- [1] Baiocco, S., Fernández-Macías, E., Rani, U. & Pesole, A. (2022). *The Algorithmic Management of Work and its Implications in Different Contexts*. Seville: European Commission, 2022, JRC129749.
- [2] Eurofound. (2018). *Automation, digitization and platforms: Implications for work and employment*. Luxembourg: Publications Office of the European Union.

- [3] Xue, C., Tian, W. & Zhao, X. (2020). The Literature Review of Platform Economy. *Scientific Programming*, 1-7. <https://doi.org/10.1155/2020/8877128>
- [4] Evans, P. C. & Gawer, A. (2016). *The Rise of the Platform Enterprise: A Global Survey*. The Center for Global Enterprise.
- [5] Stark, D. & Pais, I. (2020). Algorithmic Management in the Platform Economy. *Sociologica*, 14 (3): 47-72. <https://doi.org/10.6092/issn.1971-8853/12221>
- [6] Jarrahi, M. H., Newlands, G., Lee, M. K., Wolf, C. T., Kinder, E. & Sutherland, W. (2021). Algorithmic management in a work context. *Big Data & Society*, July-December, 1-14. <https://doi.org/10.1177/20539517211020332>
- [7] Benlian, A., Wiener, M., Cram, W. A., Krasnova, H., Maedche, A., Möhlmann, M., Recker, J. & Remus, U. (2022). Algorithmic Management Bright and Dark Sides, Practical Implications, and Research Opportunities. *Bus Inf Syst Eng* 64 (6): 825–839. <https://doi.org/10.1007/s12599-022-00764-w>
- [8] Lee, M. K., Kusbit, D., Metsky, E. & Dabbish, L. (2015). Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In: *Proceedings of the 33rd Annual ACM SIGCHI Conference*, Seoul, South Korea, 18–23 April. New York: ACM Press, 1603–1612.
- [9] Wood, A. J. (2021). *Algorithmic Management: Consequences for Work Organisation and Working Conditions*. JRC Working Papers, Series on Labour, Education and Technology, 2021/07. Seville: European Commission, EU Science Hub.
- [10] Möhlmann, M., Zalmanson, L. & Henfridsson, O. (2021). Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control. *MIS Quarterly*, 45 (4): 1999-2022. <https://doi.org/10.25300/misq/2021/15333>
- [11] Mateescu, A. & Nguyen, A. (2019). Algorithmic Management in the Workplace. *Data & Society*, February. https://datasociety.net/wp-content/uploads/2019/02/DS_Algorithmic_Management_Explainer.pdf
- [12] Rosenblat, A. & Stark, L. (2016). Algorithmic Labour and Information Asymmetries: A Case Study of Uber’s Drivers. *International Journal of Communication*, 10: 3758-3784. <http://ijoc.org/>
- [13] Duggan, J., Sherman, U., Carbery, R. & McDonnell, A. (2019). Algorithmic management and app-work in the gig economy: A research agenda for employment relations and HRM. *Human Resource Management Journal*, 30: 114-132. <https://doi.org/10.1111/1748-8583.12258>

Quarterly Trends in Soil Moisture and Air Temperature: A Python-Based Analysis

Milena Zeljković¹, Dinu Dragan¹, Srđan Popov¹ and Tanja Vranić¹

¹ University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, Novi Sad, Serbia

milenapocuca@uns.ac.rs; dinud@uns.ac.rs; srdjanpopov@uns.ac.rs; tanjanovakovic@uns.ac.rs

Abstract:

In this study, we developed a Python-based application to process and analyze soil moisture and air temperature sensor data. Utilizing Python's powerful libraries – Pandas, Matplotlib, Seaborn, Statsmodels, and Tabulate – we efficiently handled large datasets, created informative visualizations, and conducted rigorous statistical analyses. The application involved loading sensor data, preprocessing it to handle missing values and data type conversions, and then generating various plots to visualize relationships between air temperature and soil moisture. Specifically, scatter and regression plots were produced to illustrate correlations and trends. Statistical analysis using Statsmodels provided insights into the significance and strength of these relationships through Ordinary Least Squares (OLS) regression, with p-values and R-squared metrics reported. The results were systematically summarized and formatted using Tabulate for clear reporting. The application demonstrated that the relationships between air temperature and soil moisture varied across different quarterly periods, with interpretations of correlation strength and model variance provided. This approach highlights Python's versatility and effectiveness in handling, visualizing, and analyzing sensor data, making it an excellent choice for developing robust data processing applications.

Keywords:

sensor, air temperature, soil moisture, regression, python

1. Introduction

Understanding the intricate dynamics between soil moisture and air temperature is vital for effective agricultural management and environmental monitoring, especially in the context of climate variability. Recent research [1] highlights a hypersensitive relationship between soil moisture and air temperature, identifying a critical threshold known as the soil moisture breakpoint (SMbp), which occurs just above the permanent wilting point (WP). This sensitivity arises from declining latent heat flux as soil moisture decreases, significantly impacting temperature responses, emphasizing the need for climate models to accurately represent these interactions for improved predictions of extreme temperature events. Another research [2] has highlighted the significant impact of soil moisture updates on temperature forecasting within atmospheric cycling. A study conducted during May 2019 demonstrated that varying soil moisture analyses led to considerable discrepancies in temperature predictions, particularly in the lower troposphere. These findings emphasize that neglecting soil moisture updates can result in a warm bias in forecasts, while systematic updates can improve prediction accuracy, underscoring the necessity of accurate soil moisture data for effective weather forecasting. A paper [3] discusses the impact of soil moisture on climate variability, focusing on its role in the Earth system and feedback mechanisms. It highlights the importance of accurately representing soil moisture in climate models to improve predictions of weather patterns and climate extremes. The authors emphasize the need for enhanced observational data and modeling techniques to better understand soil moisture dynamics and their influence on atmospheric processes.

As part of the international IPA cross-border cooperation program between Hungary and Serbia, known as “Water Shortage Hazard and Adaptive Water Management Strategies in the Hungarian-Serbian Cross-Border Region,” a remote monitoring system was developed. This innovative system employs a wireless sensor network specifically designed for agriculture and environmental monitoring, facilitating

the continuous assessment of critical drought parameters, including soil moisture and meteorological data.

The system comprises several independent monitoring stations strategically located across various terrains in the Bačka region of the Autonomous Province of Vojvodina, Republic of Serbia. Each station is equipped with a suite of digital sensors that capture vital environmental metrics: soil moisture, air temperature, humidity, wind direction and speed, and precipitation. Notably, six analog sensors measure soil moisture at three different depths, providing comprehensive insights into water retention and usage in diverse soil. The temperature sensors operate within a broad range from -40°C to $+65^{\circ}\text{C}$, while wind sensors can accurately measure speeds up to 60 m/s. Rain gauges are incorporated to monitor precipitation, capable of measuring up to 1000 mm.

Data collected by the stations are transmitted wirelessly to a remote base station, which aggregates the information and forwards it to a centralized server. This real-time data transmission via mobile networks allows users to access current conditions and download datasets through a web application or smartphone app. The application not only presents graphical representations of monitored parameters but also offers data in multiple formats (XML, CSV, XLS) for further spatial and statistical analysis.

The establishment of this comprehensive database serves as a foundation for predicting future adverse events related to climate risks, providing a crucial resource for enhancing agricultural resilience. By leveraging soil moisture sensors, the system aims to improve understanding of water depletion rates across different soil types, thereby promoting more efficient irrigation practices.

In this paper, we present a Python-based application developed to process and analyze the sensor data, revealing trends and relationships between soil moisture and air temperature over quarterly periods. Through the utilization of Python's powerful libraries for data handling, visualization, and statistical analysis, this study elucidates the complexities of these interactions and underscores the potential of advanced data processing solutions in environmental monitoring.

2. Technology

In today's technological landscape, selecting the right tools for developing an app that processes soil moisture and air temperature sensor data can be challenging. Our optimal choice for this task was Python with its libraries matplotlib, pandas and seaborn. R programming language is known for statistical analysis and visualization with ggplot2 [4] and is strong but less flexible for integration compared to Python. Python stands out for developing applications that manage sensor data due to its comprehensive libraries.

Pandas offers DataFrames, which are flexible and efficient structures for handling large volumes of sensor data. Its features include:

- Data cleaning – Functions for handling missing values, filtering data and transforming [5] datasets,
- Data aggregation – Grouping and aggregating data for summarization and analysis,
- Time series analysis – Built-in support for time series data, crucial for sensor readings over [5] time.

Matplotlib is a foundational library for creating static, animatic, and interactive plots in Python. Its advantages include:

- Customizability – Detailed control over plot aesthetics and functionalities [6],
- Integration – Seamless integration with Pandas DataFrame and other scientific libraries,
- Versatility – Ability to create a wide range of plot including line graphs, scatter plots, histograms, and more [6].

Seaborn builds on Matplotlib and provides a higher-level interface for creating aesthetically pleasing statistical graphics. Its strengths include:

- Ease of use – Simplifies the creation of complex visualizations with minimal code [7],
- Built-in themes – Provides attractive default themes and color palettes,
- Statistical plotting – Functions for visualizing statistical relationships, which are beneficial for analyzing sensor data [7] trends.

Statsmodels is another crucial library used in this application, offering comprehensive tools for statistical modeling and hypothesis testing. Its role in the code includes:

- Regression analysis - Performing Ordinary Least Squares (OLS) regression to model the relationship between air temperature and soil moisture. This involves fitting a regression model, extracting p-values, and calculating R-squared metrics to evaluate the model's performance [8],
- Statistical Significant Testing - Providing p-values to assess the significance of predictors, which helps in understanding the impact of air temperature on soil moisture [8].

Tabulate is used for generating formatted text tables. It is employed in the code to create a well-organized summary table of the regression analysis results, making the report more readable [9].

Python's combination of Pandas, Matplotlib, Seaborn, Statsmodels and Tabulate offers a comprehensive suite for developing an application that processes and visualizes sensor data. Pandas provides robust data manipulation capabilities, Matplotlib and Seaborn enhance the visual presentation of data, and Statsmodels delivers powerful statistical analysis tools.

For an application focused on soil moisture and air temperature sensor data, Python's [10] ecosystem not only facilitates efficient data handling but also allows for sophisticated and visually appealing data presentation. This combination makes Python, along with its data science libraries, the most suitable choice for developing a powerful and user-friendly sensor data processing application.

3. Data processing

The dataset is obtained from the sensor measurements file in a CSV format. Sensor values were recorded from June 1, 2014, to April 14, 2020. The file consists of 11 sensors, of which two are the significant ones – SM1, the soil moisture sensor and AT1 – the air temperature sensor. The data was loaded into a Pandas DataFrame with ; as the delimiter. Unnecessary columns were removed to streamline the analysis. Specifically, the Uredjaj column which did not contribute to the analysis was dropped. The Vrednost column, representing the sensor readings, was converted from a string format to a float. This involved replacing commas with periods to adhere to the standard decimal notation. The Vreme column containing timestamp information, was converted to a datetime format to facilitate time-based operations. The dataset was filtered to include only relevant sensor readings, specifically those with ID vrednosti values SM1(soil moisture) and AT1(air temperature). A pivot table was created with Vreme as the index and ID vrednosti as columns. This transformation resulted in a DataFrame with columns for soil moisture and air temperature, indexed by time. The data was resampled quarterly using the resample('QE') function to analyze the seasonal trends and patterns.

Scatter plots and regression plots were generated for each quarter to visualize the relationship between soil moisture and air temperature.

Correlation coefficients and regression models were computed to assess the strength and nature of the relationships.

4. Results

Scatter plots illustrated the relationship between air temperature and soil moisture for each quarter. Each plot visualized data points with air temperature on the x-axis and soil moisture on the y-axis.

Regression plots included fitted regression lines to show the trend between air temperature and soil moisture. The lines were computed using Ordinary Least Squares (OLS) regression, and the plots were visually inspected for any discernible patterns or trends.

The correlation coefficient for each quarter was calculated to measure the strength and direction of the linear relationship between air temperature and soil moisture. The values ranged from strong negative correlations to weak or negligible correlations, depending on the quarter.

For each quarter, an OLS regression model was fitted to predict soil moisture based on air temperature. The regression results were summarized and key metrics such as R-squared, adjusted R-squared and coefficients.

At the end of the analysis, the following counts were recorded across all quarters:

1. Correlation Types:
 - Weak correlation: 10
 - Moderate correlation: 8

- Strong correlation: 6
- 2. Variance amount in the dependent variable:
 - Very low: 12 ($R\text{-squared} < 0.1$)
 - Moderate: 12 ($0.2 \leq R\text{-squared} < 0.5$)
 - High: 0 ($R\text{-squared} \geq 0.5$)
- 3. Statistical significance of predictor variable (air temperature):
 - Significant: 21 ($p\text{-value} < 0.05$)
 - Not significant: 3 ($p\text{-value} \geq 0.05$).

The detailed results, including quarterly analysis, regression summaries, and interpretations, are summarized in the following table (Table 1).

Table 1:
Report summary

Quarter	Correlation coefficient	R-squared	P-value (const)	P-value (temp)	Interpretation
Q2_2014	-0.06	0	0	0.01	Weak correlation; Very low; Significant
Q3_2014	0.37	0.14	0	0	Moderate correlation; Moderate; Significant
Q4_2014	-0.03	0	0	0.158	Weak correlation; Very low; Not significant
Q1_2015	-0.54	0.29	0	0	Strong correlation; Moderate; Significant
Q2_2015	-0.18	0.03	0	0	Weak correlation; Very low; Significant
Q3_2015	-0.46	0.21	0	0	Moderate correlation; Moderate; Significant
Q4_2015	0.34	0.12	0	0	Moderate correlation; Moderate; Significant
Q1_2016	-0.43	0.18	0	0	Moderate correlation; Moderate; Significant
Q2_2016	0.04	0	0	0.055	Weak correlation; Very low; Not significant
Q3_2016	-0.52	0.27	0	0	Strong correlation; Moderate; Significant
Q4_2016	0.5	0.25	0	0	Moderate correlation; Moderate; Significant
Q1_2017	-0.17	0.03	0	0	Weak correlation; Very low; Significant
Q2_2017	-0.04	0	0	0.089	Weak correlation; Very low; Not significant
Q3_2017	-0.58	0.34	0	0	Strong correlation; Moderate; Significant
Q4_2017	-0.18	0.03	0	0	Weak correlation; Very low; Significant
Q1_2018	-0.57	0.32	0	0	Strong correlation; Moderate; Significant
Q2_2018	-0.05	0	0	0.024	Weak correlation; Very low; Significant
Q3_2018	0.63	0.4	0	0	Strong correlation; Moderate; Significant
Q4_2018	0.14	0.02	0	0	Weak correlation; Very low; Significant
Q1_2019	0.2	0.04	0	0	Moderate correlation; Very low; Significant
Q2_2019	-0.23	0.05	0	0	Moderate correlation; Very low; Significant
Q3_2019	0.15	0.02	0	0	Weak correlation; Very low; Significant
Q4_2019	0.34	0.11	0.242	0	Moderate correlation; Moderate; Significant
Q1_2020	0.66	0.43	0.32	0	Strong correlation; Moderate; Significant

4.1. Example interpretation

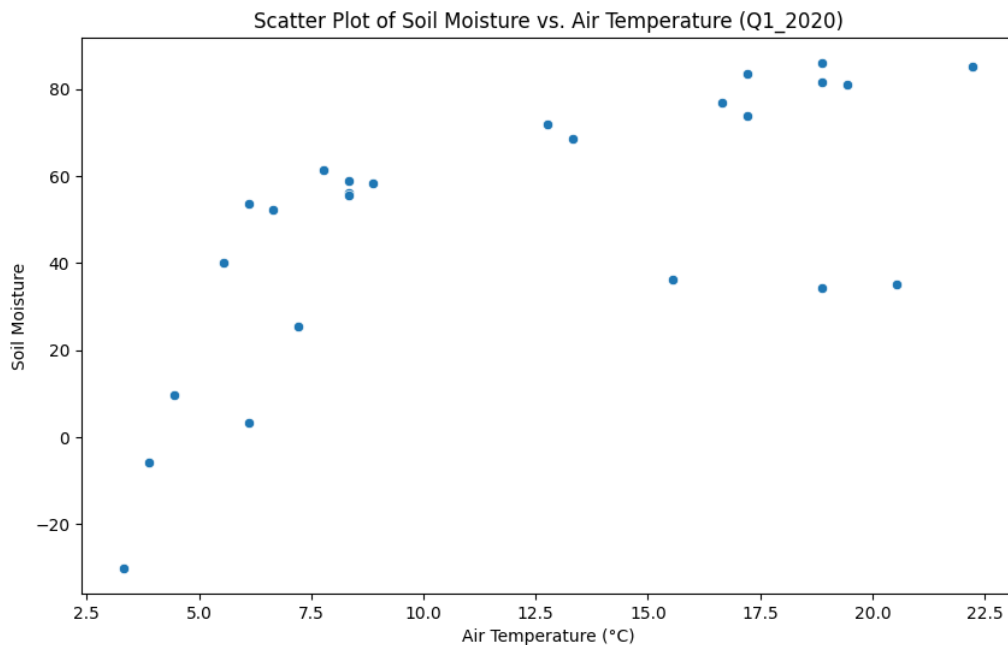


Figure 1: Scatter plot for Q1_2020

This scatter plot displays the relationship between air temperature and soil moisture. Data points were plotted with air temperature on the x-axis and soil moisture on the y-axis, showing a noticeable trend.

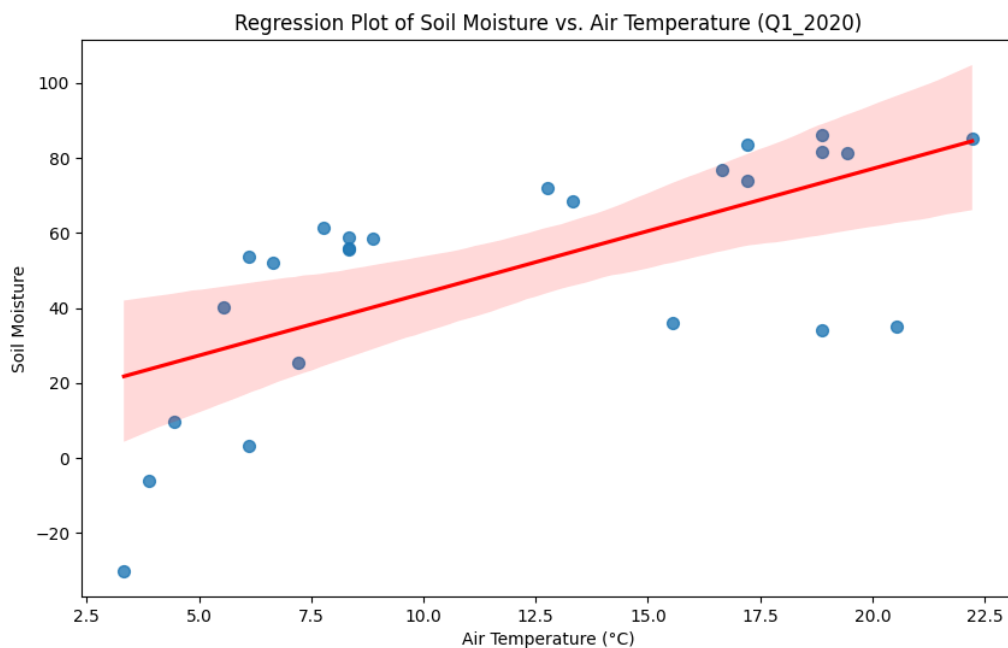


Figure 2: Regression plot for Q1_2020

The regression plot includes a fitted line from an Ordinary Least Squares (OLS) regression model. This line illustrates the relationship between air temperature and soil moisture for the quarter.

Quarter 1 (Q1_2020):

Correlation coefficient: 0.657

Regression summary for Quarter 1 (Q1_2020):

OLS Regression Results

Dep. Variable:	soil_moisture	R-squared:	0.431
Model:	OLS	Adj. R-squared:	0.407
Method:	Least Squares	F-statistic:	17.45
Date:	Tue, 17 Sep 2024	Prob (F-statistic):	0.000362
Time:	10:19:16	Log-Likelihood:	-113.65
No. Observations:	25	AIC:	231.3
Df Residuals:	23	BIC:	233.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	10.7423	10.562	1.017	0.320	-11.107	32.592
air_temperature	3.3213	0.795	4.177	0.000	1.677	4.966

Omnibus:	4.426	Durbin-Watson:	0.628
Prob(Omnibus):	0.109	Jarque-Bera (JB):	3.905
Skew:	-0.932	Prob(JB):	0.142
Kurtosis:	2.477	Cond. No.	29.6

Figure 3: Regression summary for Q1_2020

Regression Interpretation for Quarter 1 (Q1_2020):	
P-value for constant:	0.3197
P-value for air_temperature:	0.0004
Correlation coefficient:	0.657 - Strong correlation
R-squared:	0.431
The model explains a moderate amount of the variance in the dependent variable.	
The predictor variable (air_temperature) is statistically significant.	

Figure 4: Regression interpretation for Q1_2020

The correlation coefficient for Q1_2020 is 0.657 which indicates a strong positive correlation between air temperature and soil moisture, suggesting that as the air temperature increases, the soil moisture increases as well.

R-squared value is 0.431 which indicates that approximately 43.1% of the variance in the soil moisture can be explained by air temperature. This suggests a moderate level of explanation of the variability in soil moisture by air temperature.

Adjusted R-squared is 0.407 which is slightly lower than R-squared, accounts for the number of predictors in the model. This value indicated that the model still explains a significant proportion of the variance, even after adjusting for the number of observations.

P-value for constant term is 0.3197 and it is greater than the conventional significance level (0.05), indicating that the constant is not significant in predicting soil moisture.

P-value for air temperature is 0.0004, which is much less than 0.05, suggesting air temperature is a statistically significant predictor of soil moisture. The low p-value indicates strong evidence against the null hypothesis that the coefficient for air temperature is zero.

Constant term coefficient is 10.7423. It represents the intercept of the regression line. It is not statistically significant, implying that its impact on soil moisture is not substantial in this model.

The coefficient for air temperature is 3.3213 which indicates that for each one-unit increase in air temperature, soil moisture increases by approximately 3.3213 units.

The regression model shows that air temperature has a statistically significant effect on soil moisture for Q1_2020. The moderate R-squared value indicates that while air temperature explains a meaningful portion of the variance, other factors may also be at play.

It should also be pointed out that results for other quarters significantly vary. There are quarters in which the model explains a very little of the variance in the dependent variable, but the air temperature is statistically significant which is illustrated in Table 1.

5. Conclusion

This study examined the relationship between air temperature and soil moisture over various quarters, providing insights into how air temperature influences soil moisture levels. The analysis involved regression modeling and correlation assessment, summarized in a detailed table and comprehensive interpretation.

The analysis of air temperature and soil moisture reveals a complex relationship with significant variability across different quarters. The correlation between air temperature and soil moisture was categorized into weak, moderate, and strong, with a notable number of quarters exhibiting moderate to strong correlations. Specifically, air temperature showed a strong correlation with soil moisture in 6 quarters, indicating that temperature is a robust predictor in these instances.

The regression models demonstrated that air temperature explains a moderate amount of the variance in soil moisture, with 12 quarters showing moderate variance explanation and 12 quarters with very low variance explanation. This variability highlights that while air temperature is a significant factor, it is not the sole determinant of soil moisture levels.

Statistical significance was observed in 21 quarters, underscoring the reliability of air temperature as a predictor in most cases. However, there were 3 quarters where the statistical significance was not achieved, indicating that other factors may influence soil moisture during these periods.

Overall, the findings indicate that air temperature is a valuable predictor of soil moisture but that the strength of this relationship varies over time. Future research should consider additional variables such as precipitation and soil type and explore non-linear models to better capture the complexities of soil moisture dynamics.

Acknowledgment:

This research has been supported by the Ministry of Science, Technological Development, and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad" (No. 01-3394/1).

References:

- [1] Hsu, H., Dirmeyer, P., Seo, E., "Exploring the Mechanisms of the Soil Moisture-Air Temperature Hypersensitive Coupling Regime", *Water Resour. Res.* 60 (2024) e2023WR036490, doi: 10.1029/2023WR036490.
- [2] Ha, S., Sun J., "Impact of Soil Moisture Updates on Temperature Forecasting", *Geophys. Res. Lett.* 51 (2024) e2024GL110283. doi:10.1029/2024GL110283.
- [3] Thoma, R. H. M. B. R., de Vries, J. S. H. H. E. R., Deen, P. A. A. F. E. M., van der Veen, C. R. M. M., and van der Ent, E. C. P. S. "Soil moisture impacts on climate variability: Understanding

feedback mechanisms," *Earth System Dynamics* 10, no. 2 (2019): 599-616. doi:10.5194/esd-10-599-2019.

- [4] ggplot2 documentation, URL: <https://ggplot2.tidyverse.org/index.html>
- [5] Pandas documentation, URL: <https://pandas.pydata.org/docs/index.html>
- [6] Matplotlib documentation, URL: <https://matplotlib.org/stable/index.html>
- [7] Seaborn documentation, URL: <https://seaborn.pydata.org/>
- [8] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference, 92-96.
- [9] Tabulate Documentation, URL: <https://pypi.org/project/tabulate/>
- [10] Python official website, URL: <https://www.python.org/>

A Petri Net-Based Approach to Modeling Concurrency and Parallelism in Computer Architectures

Ilija Hristoski¹ and Jelena Stojanov²

¹ “St. Kliment Ohridski” University – Bitola / Faculty of Economics – Prilep, Prilepski Branitelji St. 143, 7500 Prilep, North Macedonia

² University of Novi Sad / Technical Faculty “Mihajlo Pupin” – Zrenjanin, Đure Đakovića St., 23101 Zrenjanin, Serbia

ilija.hristoski@uklo.edu.mk; jelena.stojanov@tfzr.uns.ac.rs

Abstract:

The paper focuses on concurrency and parallelism, two closely related concepts in computer architecture that deal with how tasks are executed simultaneously, though that they differ in their underlying mechanisms and objectives. Utilizing the class of Generalized Stochastic Petri Nets (GSPNs), this study proposes four distinct Petri Net-based models to capture the behavior and logic of all possible combinations of these two concepts with a single aim to clarify and explain them. Despite their straightforwardness, the resulting generic models can serve as blueprints that can be applied to simulate, design, optimize, and verify systems that rely on concurrency and parallelism, thereby enhancing both theoretical understanding and practical implementations of such systems.

Keywords:

Concurrency, parallelism, Generalized Stochastic Petri Nets (GSPNs), modeling, TimeNET

1. Introduction

The concepts of concurrency and parallelism have long histories in the development of computer science and computer architecture, since both concepts begun developing in the 1960s [1]. Concurrency was driven by the need for time-sharing and multiprogramming systems [2], while parallelism was advanced by the development of multi-threaded and multi-core processors [3]. Both paradigms have led to the development of distributed computing, supercomputers and high-performance computing systems.

In computer science, concurrency and parallelism refer to entirely distinct concepts, even though they are often used interchangeably. Rob Pike, one of the inventors of the Go programming language, in one of his excellent talks said: “*Concurrency is about DEALING with lots of things at once. Parallelism is about DOING lots of things at once.*” According to him, this distinction emphasizes that concurrency is more about the design or structure of a program, whilst parallelism is about the execution of programs, having minded that concurrency enables parallelism and makes it easy [4] [5]. Parallelism involves the use of multiple CPU cores, with each core executing a task independently. Conversely, concurrency allows a program to manage multiple tasks even on a single CPU core by switching between tasks (or threads) without necessarily completing each one before moving to the next. A program or system can exhibit characteristics of parallelism, concurrency, neither, or a combination of both [6].

A system is considered concurrent if it can handle multiple actions simultaneously or in progress at the same time [7], allowing computation to proceed without waiting for others to complete [8]. Concurrency enables a program to manage multiple tasks on a single CPU core by interleaving execution, switching between tasks without finishing each one before moving on. It involves structuring a program so tasks can progress without necessarily being executed at the same time, utilizing techniques like time-slicing. Concurrency is about handling many tasks at once and allows for components of a program or problem to be executed in any order, enhancing speed, particularly in multi-core systems [9]. Even on a single-core CPU, concurrency is achieved through rapid task

switching. It refers to the structure and scheduling of tasks, ensuring they can progress logically at the same time, regardless of actual simultaneous execution [10].

Parallelism refers to the simultaneous execution of multiple processes or threads, improving computational speed by utilizing multiple processors or CPU cores to work on different parts of a task [11]. It requires multicore or multithreaded CPUs to execute tasks or subtasks at the same time [12], and can occur at various levels, such as instruction-level and data-level parallelism [13]. A parallel system supports the execution of multiple actions simultaneously, with each core performing a task independently [7]. The essence of parallelism is about doing tasks simultaneously, requiring multiple CPU cores or processors.

Despite their long history, concurrency and parallelism remain central to modern research and practice in computer science, particularly in computer architecture, due to ongoing challenges and evolving demands. These include the rise of multi-core processors, high-performance computing, scalability, energy efficiency, real-time systems, embedded and distributed system design, cloud computing, and new programming models for AI, machine learning, and deep learning systems.

By revisiting those two paradigms, the aim of the paper is to clarify and emphasize their differences by proposing four generic, yet quite simple simulation models based on the utilization of the class of Generalized Stochastic Petri Nets (GSPNs). However, it is worth noting that beyond Petri Nets and their various extensions (e.g., Colored Petri Nets for modeling data-dependent concurrency, Timed Petri Nets for modeling real-time concurrent systems), various other formalisms, tools, and approaches have been developed to model concurrency and concurrent computing. These alternatives differ in their treatment of concurrency, synchronization, and inter-process communication. Among others, some of the most prominent alternatives include Process Calculi (Calculus of communicating systems, Communicating Sequential Processes – CSPs, π -calculus), the Actor Model, State Machines and Automata (Finite State Machines, Timed Automata), Programming Models (Thread-based Models, Message-Passing Models, and Dataflow Models), the Parallel Random-Access Machine (PRAM), Bulk Synchronous Parallel (BSP) model, as well as various software tools and programming languages, such as Simple Concurrent Object-Oriented Programming (SCOOP) and the programming languages Go (Golang) and Haskell. On the other hand, there are also alternative approaches that are used to model parallelism and parallel computing, which focus on representing, analyzing, and optimizing parallel tasks, communication, and execution in multi-core and distributed environments, such as Parallel Programming Models (Message Passing Interface – MPI, OpenMP, CUDA), Dataflow Models (Kahn Process Networks, Stream Processing), Graph-Based Models (Directed Acyclic Graphs – DAGs, Task Graphs in OpenCL), Formal Parallel Models (Parallel Random-Access Machine – PRAM, Bulk Synchronous Parallel (BSP) Model), and High-Level Parallel Languages and Libraries (Chapel, Cilk, Threading Building Blocks – TBBs).

The structure of the paper is as follows: Section 2 offers a concise overview of the class of Generalized Stochastic Petri Nets (GSPNs). In Section 3, divided into four subsections, we detail the proposed GSPN-based models that integrate both concurrency and parallelism. Section 4 presents a discussion of the proposed solutions. Finally, the concluding section summarizes the key findings and provides recommendations for future research.

2. Generalized Stochastic Petri Nets

Petri Nets are a well-known mathematical and graphical modeling tool widely used to represent and analyze concurrent, distributed, and dynamic systems. Generalized Stochastic Petri Nets (GSPNs) are an advanced class of Petri Nets, representing an extension of the class of Stochastic Petri Nets (SPNs), designed to model systems that exhibit both deterministic and stochastic behaviors. Initially introduced by Ajmone Marsan in 1984, GSPNs significantly extend the capabilities of standard (i.e., non-timed) Petri Nets by incorporating timing information, thus enabling the analysis of performance, reliability, and availability of concurrent and distributed systems [14] [15] [16] [17].

Without any intention to elaborate on the building blocks, their graphical representation, and operational behavior behind GSPNs in a more detailed manner, we hereby simply present their formal definition.

Formally, a GSPN is defined as a tuple $G = (P, T, I, O, M_0, \lambda)$, where:

- P is a finite set of places;
- $T = T_{imm} \cup T_{timed}$ is a set of transitions, with T_{imm} denoting immediate transitions and T_{timed} denoting timed transitions;
- I and O represent the input and output functions, mapping places to transitions and vice versa;
- M_0 is the initial marking of the Petri Net, representing the starting state of the system;
- $\lambda: T_{timed} \rightarrow \mathbb{R}^+$ maps each timed transition to its associated firing rate.

It is also worth noting that GSPNs can incorporate inhibitor arcs, which greatly enhance their basic functionality, despite not being included in the standard definition of GSPNs.

The stochastic nature of GSPNs allows for performance analysis through the generation of continuous-time Markov chains (CTMCs). Each reachable marking (state) of the GSPN corresponds to a state in the CTMC, and the transitions between states are governed by the firing rates of the timed transitions.

GSPNs are widely used in the modeling of complex systems, such as computer systems, computer components, communication networks, manufacturing systems, various operations, protocols, strategies, etc., where random events, probabilistic choices, and time delays are inherent.

3. GSPN-based Models of Concurrency and Parallelism

The following four subsections present and elaborate on the four possible combinations of concurrency and parallelism found in computer architectures. It is crucial to distinguish between concurrency and parallelism when determining the most effective approach for solving large-scale problems, even though these terms are often used interchangeably in practice. By focusing on how CPU cores manage tasks at the hardware level, the GSPN-based solutions provided here play a vital role in clarifying these two closely related yet fundamentally distinct paradigms.

For simplicity reasons, it is assumed that there are at most two tasks per CPU core and there are at most two CPU cores in the system, which is an absolute minimum. Moreover, in the first two cases, both dealing with the concept of non-parallelism, there are just two tasks in the system, *Task_A* and *Task_B*. In the rest of the cases that deal with the parallelism of tasks, *Task_C* and *Task_D* are added and also taken into account. It is also supposed that each of these tasks is composed of arbitrary number of subtasks (M , N , P , and R), respectively, that can be processed independently by CPU cores. As usual, the arrival rate of subtasks in a CPU core is labeled with the Greek letter λ , whilst the service rate of CPU cores is labeled with the Greek letter μ . Both the arrival times of tasks in CPU cores and CPU cores' service times are supposed to follow exponential distribution.

All hereby presented GSPN-based models are built and verified using TimeNET 4.5, a dedicated software tool suitable for modeling, analysis, and performance evaluation of stochastic systems using extended Petri Nets, including the class of Generalized Stochastic Petri Nets (GSPNs). TimeNET provides a powerful development environment suitable for the simulation and formal analysis of systems that exhibit concurrency, synchronization, and random behavior, making it particularly useful for evaluating systems where timing and probabilistic events play a key role [18] [19] [20] [21].

3.1. Case #1: Non-concurrent, non-parallel tasks

In this case, the system processes all tasks one at a time, sequentially (i.e., one after the other), by the means of a single CPU core (Figure 1).

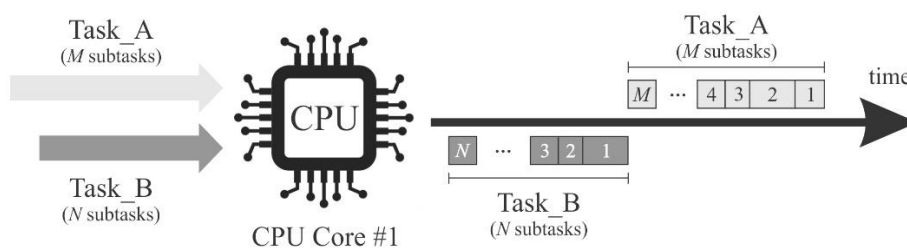


Figure 1: Non-concurrent and non-parallel execution of tasks (Source: The authors)

A single CPU core executes each task sequentially, so that *Task_A* finishes before *Task_B* begins. *Task_A* consists of M subtasks, i.e. tokens held in place P_{task_A} , while *Task_B* is comprised of N subtasks, i.e. tokens held in place P_{task_B} , as portrayed in Figure 2. The result of the execution of the GSPN model depicted in Figure 2 is equivalent to what is shown in Figure 1.

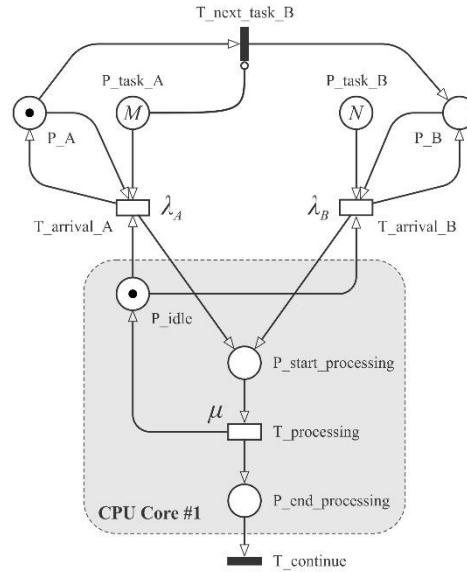
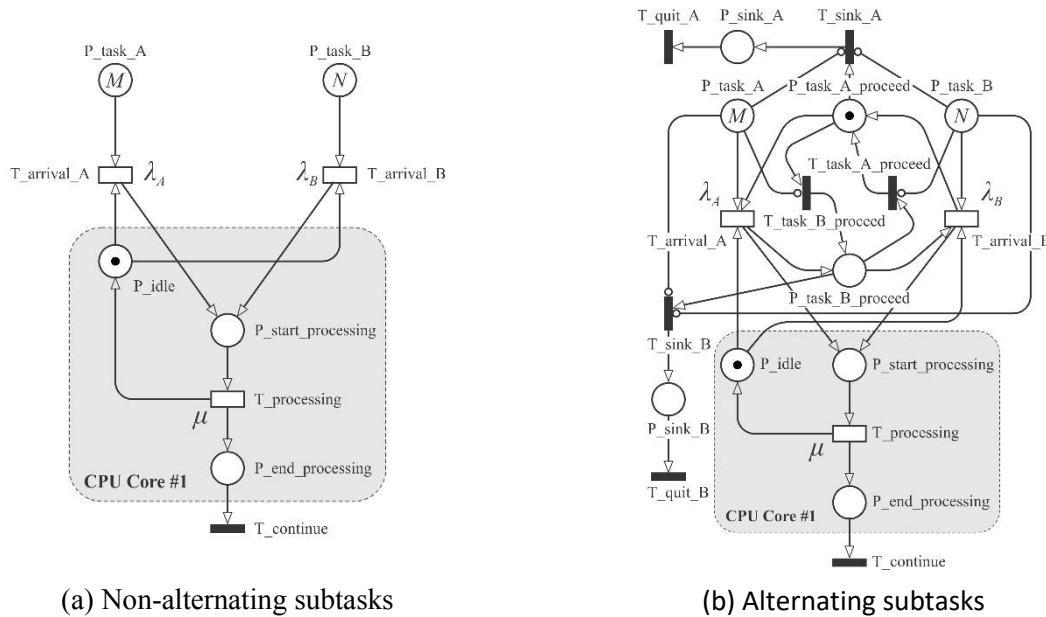


Figure 2: GSPN model resembling Case #1 (Source: The authors)

3.2. Case #2: Concurrent, non-parallel tasks

This scenario occurs when the system handles multiple tasks simultaneously, but no two tasks are executed at the exact same time. A single CPU core processes *Task_A* and *Task_B* concurrently. In the GSPN model presented in Figure 3a, the CPU core can process multiple subtasks from either *Task_A* or *Task_B*, one after the other. Meanwhile, the model in Figure 3b enforces alternating processing, where the CPU must complete the current subtask from *Task_A* before proceeding with the next subtask from *Task_B* and vice-versa, as shown in Figure 4a and Figure 4b, respectively.



(a) Non-alternating subtasks

(b) Alternating subtasks

Figure 3: GSPN model resembling Case #2 (Source: The authors)

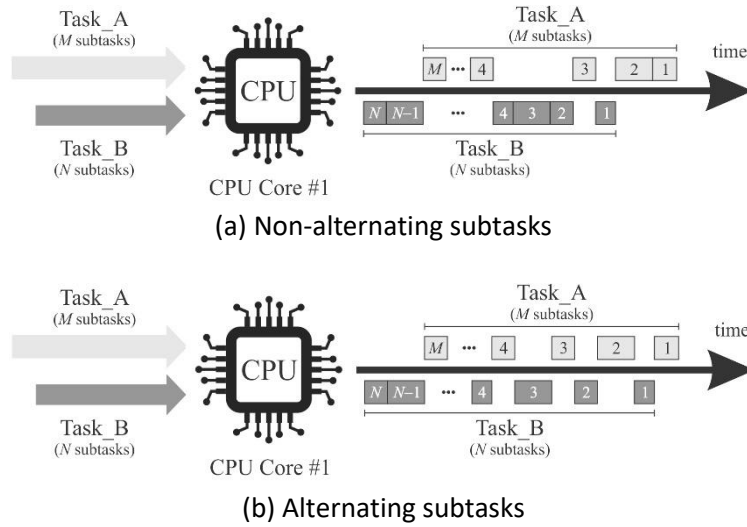


Figure 4: Concurrent and non-parallel execution of tasks (Source: The authors)

3.3. Case #3: Non-concurrent, parallel tasks

In this case, the system processes multiple subtasks of a task in multi-core CPU at the same time. In this case, two CPU cores execute each task in parallel, as shown in Figure 5. The result of the execution of the GSPN model depicted in Figure 5 is equivalent to what is shown in Figure 6.

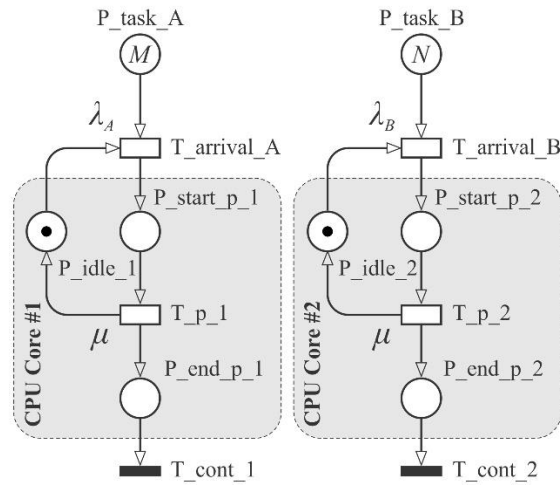


Figure 5: GSPN model resembling Case #3 (Source: The authors)

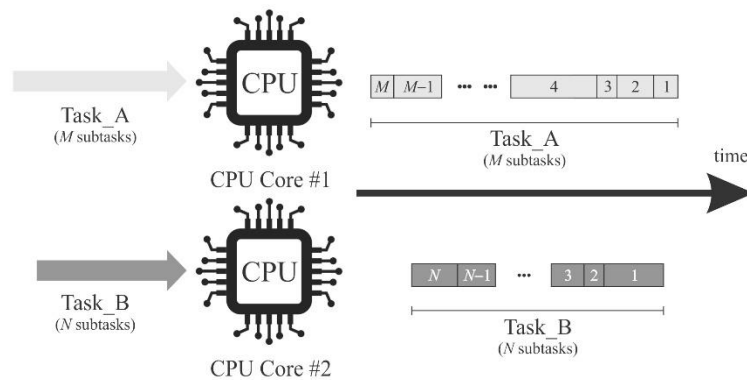


Figure 6: Non-concurrent and parallel execution of tasks (Source: The authors)

3.4. Case #4: Concurrent, parallel tasks

This is the case when the system processes multiple tasks concurrently in a multi-core CPU at the same time. In this particular case, there are two CPU cores; each of them executes two tasks concurrently; *Task_A* and *Task_B* are processed concurrently by CPU Core #1, whilst *Task_C* and *Task_D* are processed concurrently by CPU Core #2. At the same time, both *Task_A* and *Task_B* are processed in parallel with *Task_C* and *Task_D* (Figure 7).

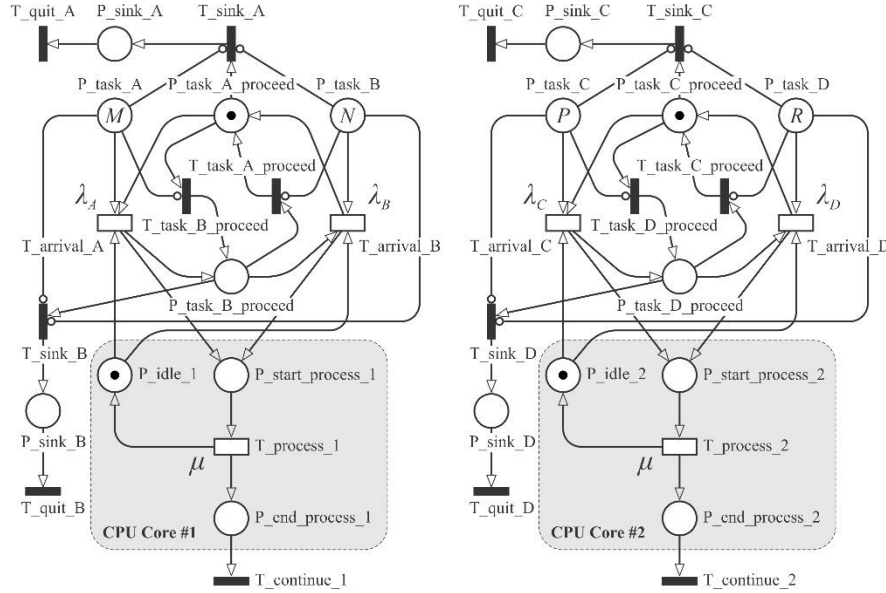


Figure 7: GSPN model resembling Case #4 (Source: The authors)

The result of the execution of the GSPN model depicted in Figure 7 is equivalent to what is shown in Figure 8.

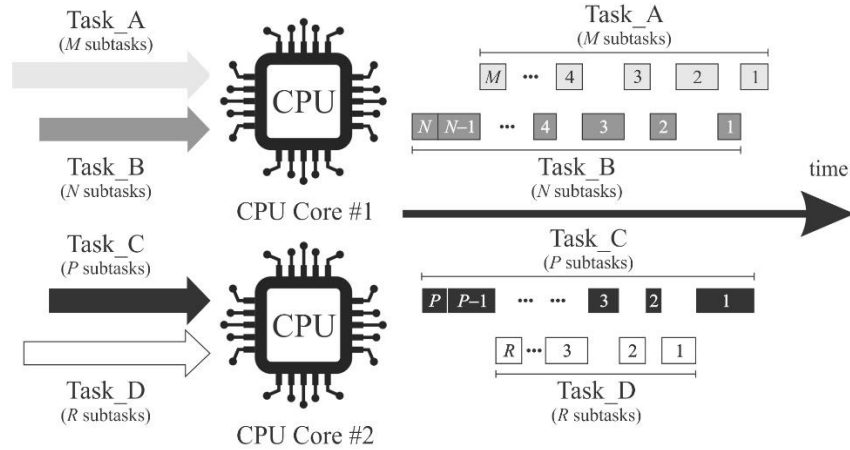


Figure 8: Concurrent and parallel execution of tasks (Source: The authors)

4. Discussion

Despite the fact that GSPNs are a powerful tool for modeling and analyzing systems involving concurrency and parallelism, they face several *limitations* when applied to real-world systems. The most significant limitation is the state space explosion problem, which occurs when the complexity of the modeled system increases by including more concurrent processes (i.e., tasks), larger number of tokens, or transitions. Further on, a serious limitation of this approach can be addressing complexity

intrinsic to modeling real-world systems where the processing of tasks involve non-Markovian time distributions, such as deterministic, fixed delays or other types of time distributions that are not exponential. Yet another important limitation is the difficulty of modeling complex synchronization mechanisms frequently found in many computer architectures that control the access of multiple tasks (i.e., threads or processes) to shared resources, which might require complex and potentially unwieldy Petri Net structures, making the model harder to interpret and analyze, even computationally intractable. At last but not at least, there is always the limitation known as the trade-off between abstraction level and scalability: GSPNs provide a high level of abstraction, which is advantageous for capturing general behaviors of systems, but this feature can become a limitation when low-level details are important. The need to model such intricate details can make GSPN models either too abstract, which can lead to losing important information, or too complex by reducing the simplicity and elegance of the Petri Net approach, thus making the analysis difficult.

The two important processes used to ensure that the model accurately represents the system being studied and behaves as expected are the processes of model verification and validation.

The *verification* of the proposed GSPN models, which refers to the process of ensuring that the GSPN model is correctly implemented according to its formal specification, has been carried out using the TimeNET's Token Game module. It proved that the proposed models are built correctly in terms of their structure, behavior, and conformance to the GSPN rules, by addressing key aspects such as structural correctness, deadlocks and liveness, boundedness, reachability analysis, and internal consistency.

On the other hand, the *validation* of these models, which refers to the process of ensuring that the GSPN model accurately represents the real-world system or process it is intended to simulate, has not been carried out yet due to the lack of real-world data, physical resources (CPUs), as well as suitable tools and measurement methods to convey effective validation. However, in this particular case, this aspect may be considered less crucial, since the proposed models are intended to be used mostly as conceptual designs, having minded their exploratory and theoretical focus.

5. Conclusions

This paper has presented a Petri Net-based approach to modeling concurrency and parallelism in computer architectures, with a focus on the use of Generalized Stochastic Petri Nets (GSPNs). The proposed formal GSPN framework, despite the ultimate straightforwardness of the presented models, offers a flexible and powerful tool for modeling key behaviors in concurrent and parallel execution of tasks in computer systems, with a huge potential in providing insights into performance, resource utilization, and potential bottlenecks. However, despite their modeling strengths, GSPNs come with several limitations that affect their applicability in large-scale, real-world systems, as it was elaborated in the previous section. Despite these limitations, GSPNs remain a valuable tool for quantitative analysis of systems exhibiting features of concurrency and parallelism, and can be applied to obtain metrics such as throughput, utilization, and response time.

The proposed GSPN-based models presented in this paper can be used for simulating and analyzing system performance in computer architecture and related fields, thus helping to assess how different task-handling strategies can affect resource utilization and responsiveness under various conditions, including the number of subtasks, their arrival rates, and CPU processing speeds. They can also help in designing and optimizing hardware or software systems by exploring efficient implementations of concurrency and parallelism, optimizing factors like execution speed and energy consumption. These models can also be used for verification of system behavior, and detecting issues such as deadlocks or bottlenecks. Furthermore, they can allow for the comparison of concurrency and parallelism strategies, guiding decision-makers in choosing the best approach. Ultimately, they can provide a theoretical foundation for studying and teaching these concepts and assist in resource allocation and task scheduling in multi-core or distributed systems for greater efficiency.

Future work could focus on the performance evaluation of the proposed GSPN-based models and their validation vis-à-vis real-world systems. This will ensure that GSPNs remain a practical and effective tool in the ongoing efforts to improve the design and analysis of concurrent and parallel computing systems.

References:

- [1] L. Proença, A brief history of modern computers, multitasking and operating systems, 2022. URL: <https://dev.to/leandronsp/a-brief-history-of-modern-computers-multitasking-and-operating-systems-2cbn>
- [2] L. Lamport, The Computer Science of Concurrency: The Early Years, In: Concurrency: The Works of Leslie Lamport (2019) 13–26. doi: 10.1145/3335772.3335775
- [3] A. Grama, A. Gupta, G. Karypis, V. Kumar, Introduction to Parallel Computing, 2nd. ed., Addison-Wesley, Boston, MA, USA, 2003.
- [4] R. Pike, Concurrency is Not Parallelism, Speech at Heroku Waza 2012, Video, 2012. URL: <https://www.youtube.com/watch?v=oV9rvDIIKEg>
- [5] M. Streeter, Concurrency is Not Parallelism, Rob Spike's Speech at Heroku Waza 2012, Video, 2020. URL: <https://www.youtube.com/watch?v=qmg1CF3gZQ0>
- [6] S. Marlow, Parallel and Concurrent Programming in Haskell: Techniques for Multicore and Multithreaded Programming, 1st. ed., O'Reilly Media, Sebastopol, CA, USA, 2013.
- [7] C. Breshears, The Art of Concurrency: A Thread Monkey's Guide to Writing Parallel Applications, 1st. ed., O'Reilly Media, Sebastopol, CA, USA, 2009.
- [8] A. Silberschatz, P. B. Galvin, G. Gagne, Operating System Concepts, 9th. ed., Wiley, Hoboken, NJ, USA, 2012.
- [9] L. Lamport, Time, Clocks and the Ordering of Events in a Distributed System, Communications of the ACM 21(7) (1978), 558–565. doi: 10.1145/359545.359563
- [10] P. B. Hansen (ed.), The Origin of Concurrent Programming: From Semaphores to Remote Procedure Calls, Springer-Verlag, New York, NY, USA, 2002.
- [11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, Introduction to Algorithms, 3rd. ed., MIT Press, Cambridge, MA, USA, 2009.
- [12] A. C. Sodan, J. Machina, A. Deshmeh, K. Macnaughton, B. Esbaugh, Parallelism via Multithreaded and Multicore CPUs, Computer 43(3) (2010), 24–32. doi: 10.1109/MC.2010.75
- [13] D. A. Patterson, J. L. Hennessy, Computer Organization and Design: The Hardware/Software Interface, 5th. ed., Morgan Kaufmann, Burlington, MA, USA, 2013.
- [14] M. Ajmone Marsan, G. Conte, G. Balbo, A Class of Generalized Stochastic Petri Nets for the Performance Evaluation of Multiprocessor Systems, ACM Transactions on Computer Systems 2(2) (1984), 93–122. doi: 10.1145/190.191
- [15] G. Chiola, M. Ajmone Marsan, G. Balbo, G. Conte, Generalized Stochastic Petri Nets: A Definition at the Net Level and Its Implications, IEEE Transactions on Software Engineering 19(2) (1993), 89–107. doi: 10.1109/32.214828
- [16] M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli, G. Franceschinis, Modelling with Generalized Stochastic Petri Nets. Wiley Series in Parallel Computing, John Wiley and Sons, West Sussex, UK, 1995.
- [17] G. Balbo, Introduction to Generalized Stochastic Petri Nets, in: M. Bernardo, J. Hillston (Eds.), Formal Methods for Performance Evaluation (SFM 2007), Volume 4486 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2007, pp. 83–131. doi: 10.1007/978-3-540-72522-0_3
- [18] A. Zimmermann, J. Freiheit, R. German, G. Hommel, Petri Net Modelling and Performability Evaluation with TimeNET 3.0, in: B. R. Haverkort, H. C. Bohnenkamp, C. U. Smith (Eds.) Computer Performance Evaluation: Modelling Techniques and Tools (TOOLS 2000), Volume 1786 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2000, pp. 188–202.
- [19] A. Zimmermann, M. Knoke, A. Huck, G. Hommel, Towards Version 4.0 of TimeNET, in: Proceedings of the 13th GI/ITG Conference on Measurement, Modeling, and Evaluation of Computer and Communication Systems (MMB 2006), Nurnberg, Germany, pp. 477–480, 2006.
- [20] A. Zimmermann, M. Knoke, TimeNET 4.0: A Software Tool for the Performability Evaluation with Stochastic and Colored Petri Nets – User Manual, Faculty of EE&CS Technical Report 2007-13, Technische Universität Berlin, Berlin, Germany, 2007.
- [21] A. Zimmermann, Modelling and Performance Evaluation with TimeNET 4.4, in: N. Bertrand, L. Bortolussi, (Eds.), Quantitative Evaluation of Systems (QEST 2017), Volume 10503 of Lecture Notes in Computer Science, Springer, Cham, Switzerland, 2017, pp. 300–303.

Comparative Analysis: Web, Native, and Progressive Web Applications

Sara Poparić¹, Dinu Dragan¹, Dušan B. Gajić¹, Veljko B. Petrović¹

¹ University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

poparic.sara@uns.ac.rs; dinud@uns.ac.rs; dusan.gajic@uns.ac.rs; pveljko@uns.ac.rs

Abstract:

Mobile applications are becoming an important part of everyday life. The choice among different development approaches for mobile applications is thus very important for developers and organizations. These development approaches are web, native, and progressive web application (PWA). Each approach has distinct advantages and limitations that affect performance, development cost, and user experience. This study provides a comprehensive comparison of these approaches by implementing the same application for an equestrian federation using all three of them. The analysis of the development costs focused on the codebase and its modularity, cross-platform compatibility, and maintenance and requirements. Performance tests evaluated responsive time, time to first interaction, and rendering speed. User experience tests involved ten participants evaluating the task competition across platforms and user preferences across platforms. Results showed that native applications excel in response time and device integration, while web and PWA applications provide faster interaction and rendering. Most participants preferred the native application for its smooth performance. These findings suggest that native applications are ideal for performance-critical tasks, while web and PWAs offer advantages in cross-platform accessibility and easier maintenance.

Keywords:

Web application, native application, progressive web application, performance, user experience

1. Introduction

Mobile phone applications are a key part of everyday life, used by millions for communication, entertainment, and business. With the rapid growth of mobile devices and demands on user experience, the decision on how to build these applications has become increasingly important.

Developers must choose between web applications, native apps, or progressive web applications (PWAs). Web applications offer accessibility through browsers but often lack performance [1]. Native applications provide superior performance and offline functionality, but at higher development costs [2]. PWAs aim to balance these by offering offline capabilities and easy installation, while maintaining web development simplicity [3].

Managing data across platforms demands both high performance and ease of use. The architecture choice impacts both development and user experience. This study compares development costs, performance, and user experience across web, native, and PWA implementations of the same application. The software example analyzed in this paper was previously developed to support equestrian federation.

Most previous studies focus on specific aspects, such as energy consumption, performance, or user experience [4] [5]. The key difference in this study is that it offers a comprehensive comparison of multiple aspects across the three implementations of the same application. Performance tests measured response time, time to first interactive, and rendering performance, conducted five times on three different devices. User experience evaluations involved ten participants, experienced in web and mobile applications, performing tasks in each implementation, distinguishing them, and providing feedback on their preferred application.

Results showed that the native application loaded the fastest, but web and PWA applications offered quicker interactions after loading. While the native application excels in performance, web

and PWA provide fast interaction and easier cross-platform maintenance. User experience tests revealed that most participants preferred the native app due to its smooth performance, despite the accessibility and convenience offered by web and PWA apps.

The remainder of the paper is organized as follows: Section 2 covers key differences and related work. Section 3 introduces the comparative analysis from previous work. Section 4 describes the solution, methodologies and its results. Section 5 concludes the paper and discusses future work.

2. Background

Web applications are hosted on remote servers and accessed via web browsers. Their platform independence allows them to run on any device or platform without installation. However, they require a stable internet connection, and their performance depends on connection quality [1].

Native mobile applications are built for specific platforms, such as Android or iOS, using platform-specific languages like Java or Kotlin for Android and Swift or Objective-C for iOS. These applications require installation through app stores. They can integrate deeply with device hardware and can function offline [2].

Progressive Web Applications (PWAs) offer a native-like experience while being built using standard web technologies. They can be installed directly from the browser and accessed from a device's home screen or desktop. Despite the installation, they run in a browser environment and rely on it to function. PWAs support offline use via service workers. They must be served over HTTPS to ensure a secure connection [3].

While web, native, and progressive web applications share similarities, they differ significantly in terms of installation, updates, access, and discoverability. The main differences between these implementations are summarized in Table 1 [5].

Table 1:

The main differences between web, native and PWA applications

Feature	Web Application	Native Application	PWA
Installation	Not required.	Mandatory (via app store).	Optional (via browser).
Updates	Instant, automatic updates.	Delayed - must be submitted to the app store and downloaded by users.	Instant, automatic updates.
Access via Home Screen	Not supported.	Fully supported.	Fully supported.
Offline Access	Not supported.	Fully available.	Available through caching mechanisms.
Discoverability	Easily discoverable via browser's search engines.	Restricted to app store searches.	Easily discoverable via search engines.

Previous studies have primarily focused on specific aspects such as energy efficiency, cross-platform capabilities, and user experience [4] [5]. However, they often lack a holistic comparison of development cost, performance, and user satisfaction across multiple implementations of the same application. This study addresses that gap by providing a comprehensive evaluation of web, native, and PWA implementations of the same application, offering a more complete perspective on the trade-offs involved in selecting different technology stacks. By examining development cost, performance, and user experience, this research provides valuable insights for developers and organizations in making informed decisions about technology choices for cross-platform development.

3. Comparative Analysis from Previous Work

The aim of this section is to evaluate the differences and similarities between the implementations of the web, native and progressive web application across various technical and usability metrics. The analysis will focus on development costs.

3.1. Development Costs

This section explores the key challenges encountered during the development of the web, native, and PWA applications. It focuses on codebase structure and modularity, cross-platform compatibility, and the complexities of maintenance and updates across the different platforms.

A well-organized and modular codebase is crucial throughout the software development life cycle because it affects how efficient maintenance is, as well as the collaboration between team members [14]. The analysis of how the technologies used in this study promote code reusability and assessment of code structure in terms of maintainability for future developers is shown in Table 2.

Table 2:
Codebase comparison

Application	Codebase
Web Application	React JS's component-based architecture promotes code reusability, fast development, and easier maintenance.
Native Application	React Native's component-based architecture allows code reuse between web and native apps.
PWA	Same as for the web application.

Cross-platform compatibility enables an application to run smoothly on various devices and platforms. Additionally, a single codebase reduces development time and maintenance efforts [15]. Cross-platform capabilities of each implementation are presented in Table 3.

Table 3:
Cross-platform compatibility comparison

Application	Cross-Platform Compatibility
Web Application	Runs on any browser, making it inherently cross-platform, but lacks access to device-specific features like GPS or push notifications.
Native Application	Cross-platform functionality with one codebase for Android and iOS, offering access to native device features.
PWA	Works across desktop and mobile platforms, but browser support, especially on iOS, may limit functionality.

Table 4:
Maintenance and updates comparison

Application	Maintenance and Updates
Web Application	Updates are delivered automatically, ensuring all users are on the latest version.
Native Application	Updates must be manually downloaded from app stores, with approval processes complicating the release cycle.
PWA	Updates are delivered automatically, but managing service workers adds complexity to ensuring smooth updates.

Maintenance and updates are essential to ensure bug fixes and new versions of the application reach users, and have an impact on the longevity of the application and the satisfaction of its users [16]. The maintenance strategies and update processes for each of the implemented applications are shown in Table 4.

4. Empirical Study

This section evaluates the practical performance and user experience of web, native, and PWA applications in the context of an equestrian federation. By measuring key performance metrics and collecting user feedback, this study aims to provide insights into the real-world implications of each platform's strengths and limitations.

4.1. Implemented Software Solution

The foundation of this study is a web application built using ReactJS [6]. From this base, both the native and PWA versions of the application were derived. The native application was built using the React Native framework [7]. The PWA was created by adding PWA-specific features, such as service workers and a web manifest, to the web application, enabling native-like functionalities [8] [9].

The main goal was to maintain a consistent user interface across all three implementations, allowing their differences to be easily identified. Therefore, TailwindCSS was used for styling both the web and PWA versions, offering responsive and reusable designs [10]. For the native application, Nativewind was used, which allows the reuse of TailwindCSS styles in React Native [11]. All versions of implemented software are shown in Figure 1.

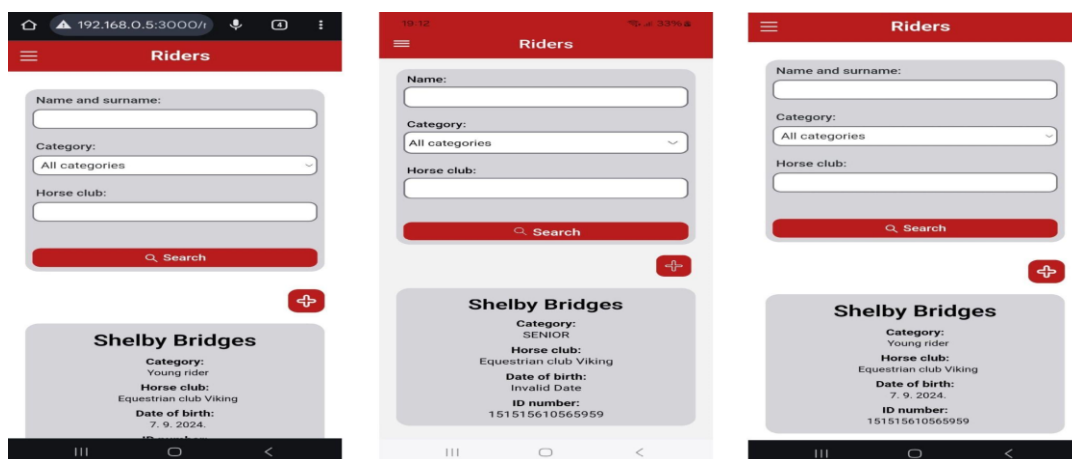


Figure 1: Web application, native application and PWA

All three applications share the same backend, which handles business logic and data management, developed using Java Spring Framework [12]. The backend interacts with a MySQL database to store and manage the application's data [13]. This centralized architecture ensures consistent data handling and synchronization across all platforms. System architecture is presented in Figure 2.

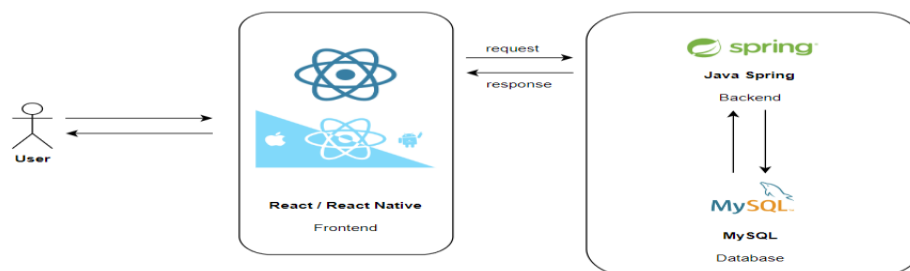


Figure 2: Application architecture

The base web application used in this study represents a part of the larger system designed for the equestrian federation. Its primary goal is to assist the federation in efficiently managing and keeping track of clubs, riders, horses, federation officials, and competitions. The system is designed to streamline the administrative processes within the federation, ensuring accurate and up-to-date information available across various platforms and devices.

The part of the application used in this study includes following functionalities:

1. Club data administration:
 - Filter clubs: Users can filter clubs by entering a club's name and/or city in which club is located.
 - Club registration: Users can register a new club in the system by entering its name and address.
2. Rider data administration:
 - Filter riders: Users can filter riders by entering a rider's name, category, and/or club.
 - Rider registration: Users can register a new rider in the system by entering its name, last name, id number, date of birth and name of the club that rider is competing for.
3. Horse data administration:
 - Filter horses: Users can filter horses by entering a horse's name, sex, and/or club.
 - Horse registration: Users can register a new horse in the system by entering its name, sex, year of birth and name of the club that rider is competing for.

4.2. Methodology

This section outlines the methodologies used for both performance testing and the questionnaire to compare web, native, and PWA applications. The methodologies aim to evaluate both the technical performance and user experience of the three types of applications.

4.2.1. Performance Testing Methodology

Performance is a key factor in the success of any application, as it directly influences user experience [17]. This section compares the performance of web, native, and progressive web applications by measuring three key metrics: response time, time to first interaction (TTFI), and rendering performance.

Response Time was measured as the time taken for the screen to start loading after initiating navigation (navigation from 'Clubs' to 'Riders' screen) [18]. Time to First Interaction (TTFI) was calculated from when a 'Riders' page started loading until all data was fully rendered and the app became interactive [20]. Rendering Performance was assessed using Chrome DevTools for web and PWA. For the native app, the react-native-performance library is measured from the start of rendering to the onLayout event, which signals that the root component has been rendered and laid down. Rendering and painting times were summed to provide a comprehensive measure [21].

The tests were conducted five times on three different devices: Samsung A52, Samsung S23+, and Samsung A10, to ensure consistency and account for hardware differences. All tests were performed under consistent network conditions to ensure that network variability did not influence the results, especially for web and PWA applications.

4.2.2. Questionnaire Methodology

An ideal user experience in mobile applications would deal with a user-centered, intuitive, accessible, and efficient design for better user engagement and retention of users' interest in the application [22]. This section outlines the methods that are used to test the user experience across the web, native, and progressive web applications developed for this study. Through a series of tasks and questions, the idea is to gauge how users would interact with each type of application, their preferences, and their ability to distinguish between these implementations.

The questionnaire included 10 male participants all between 20 and 35 years old. All participants are IT professionals. This choice ensured that participants had the necessary background to understand and evaluate the differences between web, native, and progressive web applications. It was conducted through online form. Each participant completed the questionnaire individually, ensuring they had equal exposure to each app type. The questionnaire was administered at the 7th of September 2024, ensuring consistency in the conditions under which feedback was collected.

The questionnaire included the following sections:

- **App Identification Task:** Participants were asked to identify the type of app (web, native, PWA) based on screenshots from Figure 1 and given explanation.
- **Task Performance Assessment:** Participants were asked to complete a rider registration form in each of three applications. The form required the following fields: the rider's name, last name, ID number, date of birth, age category and horse club. The date of birth, age category, and horse club fields were chosen using select inputs (dropdown menus). Then, they had been asked to rate the ease of completion on a scale from 1 to 5.
- **User Preferences:** Participants selected their preferred app type and explained their choice.

4.3. Results and Discussion

Table 5:

Results of performance testing of the web, native and PWA application

	Web Application	PWA	Native Application
Response Time	26ms	24ms	6ms
TTFI	35ms	32ms	243ms
Rendering Time	17ms	16ms	25ms

4.3.1. Performance Testing Results

The web and PWA apps had consistent response times, while the native app showed a different pattern. The first navigation took 6 milliseconds, but subsequent navigations were nearly instantaneous (around 0ms). This is common in native apps due to in-memory caching, where previously loaded screens are kept in memory, eliminating the need for re-rendering or re-fetching data [19].

The native application had a significantly longer TTFI compared to the web and PWA apps. This is likely due to the overhead of rendering native components, initializing native modules, and JavaScript-native communication. Native apps typically require more setup time, while web and PWA apps benefit from lighter initial processes due to caching and preloading [19].

Minimal differences in rendering performance were observed. The native app took slightly longer, likely due to React Native's rendering process [19], but the delay was small enough to be unnoticeable in normal use. All three platforms provided fast and responsive rendering times.

4.3.2. Questionnaire Results

The results of App Identification Test are:

- **Native Application:** 70% identified it correctly, while 30% mistook it for a PWA, suggesting some difficulty in distinguishing between native and PWA experiences.
- **Web Application:** 100% correctly identified the web app, likely due to the recognizable browser UI.
- **PWA:** 60% identified it correctly, while 40% thought it was a native app, showing how closely PWAs mimic native app experiences.

These results suggest that while web apps are easily recognizable, distinguishing between native and PWA apps is more challenging due to their similar user interfaces, especially regarding offline capabilities.

The observed results for Task Completion Assessment are shown in Table 6. The form-completion task demonstrated a difference in how each of the implementations handled complex input, such as dropdown menus. The native app behaved the best and was rated highest on both ease of use and efficiency, probably because it had better optimization for touch. The web application performed well but was a little slower, especially when working with dropdowns. It might be related to network latency or the performance of browsers on mobile. Although the PWA was very similar in design and functionality to the native app, minor delays proved the point that network dependence does play a certain role in users' experience.

Table 6:

Average participant ratings for ease of use and response time across the different applications

	Web Application	Native Application	PWA
Ease of use	4.9	5	4.7
App response	4.3	5	4.3

The results of User Preference Test are shown in Figure 3. A significant proportion of participants (70%) expressed preference for the native application, suggesting that it offered a superior overall experience regarding user familiarity and usability. Although the web application demonstrated good performance in certain tasks, only 20% of the participants indicated a preference for it, which may be attributed to the perceived constraints associated with browser-based interactions when compared with native capabilities. Whereas the PWA has been really similar to the native app, just 10% of participants went for it, which means that users still will prefer fully native when given both options.

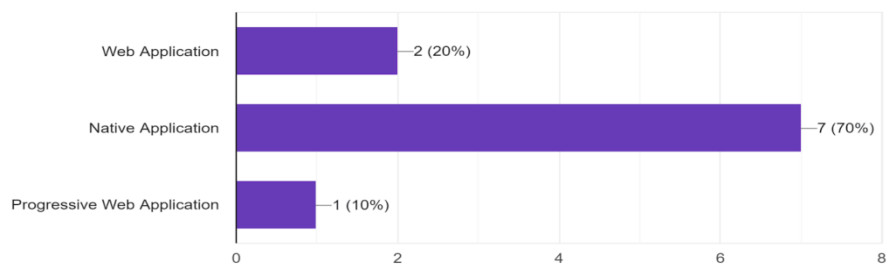


Figure 3: Participants preference on use what application they would use

5. Conclusion

This paper compares web, native, and PWA applications in the context of an equestrian federation, aiming to find out which one provides the best balance between development efficiency, user satisfaction, and performance in today's mobile-driven world.

The analysis revealed that native applications excel in performance and device integration, demonstrating significantly faster response times, particularly after the first load. In contrast, web and PWA implementations performed better in terms of time to first interaction and rendering speed, highlighting their advantages in quickly becoming usable. These theoretical strengths informed the design of the empirical tests and helped shape expectations for performance and user experience.

User experience tests showed a preference for the native app due to its smoother performance, even though web and PWAs offered accessibility and convenience across devices. These findings suggest that native apps may be more suitable for scenarios requiring fast, seamless experiences, while web and PWAs provide benefits for cross-platform deployment and maintenance.

No platform emerged as universally superior. The choice should be guided by specific project requirements, such as speed, device integration, or development constraints. Future work should explore different use cases with distinct performance needs, long-term user engagement, and battery consumption across platforms, as evolving technologies may influence the best choice of platform.

Acknowledgment:

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad” (No. 01-3394/1).

References:

- [1] iTrobes, What are the Advantages and Disadvantages of Web Applications?, URL: <https://www.itrobes.com/what-are-the-advantages-and-disadvantages-of-web-applications/>.
- [2] Margaret Rouse, Native Mobile App, 2024. URL: <https://www.techopedia.com/definition/27568/native-mobile-app>.
- [3] Meryl Kuwatani, What is a PWA?, 2023. URL: <https://onesignal.com/blog/what-is-a-pwa/>.
- [4] S. Huber, L. Demetz, PWA vs the Others: A Comparative Study on the UI Energy-efficiency of Progressive Web Apps, France, 2021, doi: 10.1007/978-3-030-74296-6-35.
- [5] S. S. Tandel, A. Jamadar, Impact of Progressive Web Apps on Web App Development, Maharashtra, India, 2018, doi: 10.15680/IJIRSET.2018.0709021.
- [6] React, A JavaScript library for building user interfaces, URL: <https://legacy.reactjs.org/>.
- [7] React Native, React Native Documentation, URL: <https://reactnative.dev/docs/getting-started>.
- [8] mdn web docs, Service Worker API, URL: https://developer.mozilla.org/en-US/docs/Web/API/Service_Worker_API.
- [9] mdn web docs, Web app manifests, URL: <https://developer.mozilla.org/en-US/docs/Web/Manifest>.
- [10] Anna Fitzgerald, Tailwind CSS: What It Is, Why Use It & Examples, 2022, URL: <https://blog.hubspot.com/website/what-is-tailwind-css>.
- [11] NativeWind, Overview, URL: <https://www.nativewind.dev/overview/>.
- [12] IBM, What Is Java Spring Boot?, URL: <https://www.ibm.com/topics/java-spring-boot>.
- [13] Jeffrey Erickson, MySQL: Understanding What It Is and How It's Used, 2024, URL: <https://www.oracle.com/mysql/what-is-mysql/>.
- [14] Ana Crudu, MoldStud Research Team, The importance of code maintainability in software projects, 2024, URL: <https://moldstud.com/articles/p-the-importance-of-code-maintainability-in-software-projects>.
- [15] Grady Andersen, MoldStud Research Team, Full Stack Development: Overcoming Obstacles in Cross-Platform Compatibility, 2024, URL: <https://moldstud.com/articles/p-full-stack-development-overcoming-obstacles-in-cross-platform-compatibility>.
- [16] Ana Crudu, MoldStud Research Team, Strategies for app maintenance and updates, 2024, URL: <https://moldstud.com/articles/p-strategies-for-app-maintenance-and-updates>.
- [17] Umut Can Koruyan, The Importance of Mobile App Performance, 2024, URL: <https://virgosol.com/en/blog/detail/the-importance-of-mobile-app-performance>.
- [18] SolarWinds, Page Load Time vs. Response Time – What Is the Difference?, 2024, URL: <https://www.pingdom.com/blog/page-load-time-vs-response-time-what-is-the-difference/>.
- [19] Aman Dwivedi, Caching in React Native, 2023, URL: <https://www.scaler.com/topics/react-native/caching-in-react-native/>.
- [20] Philip Walton, Time to Interactive (TTI), 2023, URL: <https://web.dev/articles/tti>.
- [21] Paul Lewis, Rendering performance, 2023, URL: <https://web.dev/articles/rendering-performance>.
- [22] Builder.ai, App user experience (UX) definition: Benefits, key components & more, URL: <https://www.builder.ai/glossary/user-experience>.

Testing the capabilities of artificial intelligence and prompt engineering in the field of agrochemical soil analysis with fertilization recommendations

Željko Eremić¹ and Milana Drašković¹

¹ Technical College of Applied Sciences in Zrenjanin, Djordja Stratimirovica 23, Zrenjanin, Serbia

zeljko.eremic@vts-zr.edu.rs; draskovic.milana@gmail.com

Abstract:

The basic idea in this paper is to test the possibilities offered by artificial intelligence in the field of agrochemical soil analysis with fertilization recommendations. In doing so, the principles of prompt engineering are used in order to improve the automation of the entire process. An experiment is being conducted, the basis of which is the GPT-4o model, which we use in combination with prompt engineering principles. The input data for our experiment is an image containing the Soil analysis results table. Then, data related to fertilization recommendations from GPT-4o were requested using the advantages offered by prompt engineering. At the output, the results are obtained in JSON and HTML format. The results are discussed and compared with those given in a previous article by GPT-4. At the end, conclusions are given, from which it can be seen that GPT-4o, together with prompt engineering, has added important possibilities for such analyses. Neither GPT-4 nor GPT-4o showed a high degree of reliability in the field of agrochemical soil analysis with fertilization recommendations.

Keywords:

GPT-4o, Prompt Engineering, Internet technologies, Agrochemical soil analysis, Fertilization

1. Introduction

Although artificial intelligence in various forms has been present for decades, it has only recently come into the focus of the wider public. The company OpenAI is most responsible for this popularity. She presented ChatGPT (Chat Generative Pre-trained Transformer) at the end of November 2022. Since then, it has become very easy to interact between humans and artificial intelligence through the common web browser. A less used possibility that has emerged is the use of APIs. The API makes it possible, for example, to communicate with GPT through scripts. Prompt engineering allows us to design inputs for artificial intelligence in order to obtain outputs that are optimal for us.

At the beginning, some similar previous researches will be mentioned. We present the theoretically most important facts related to GPT-4o, prompt engineering and agrochemical soil analysis with fertilization recommendations. In our previous article [1] we tested the possibilities that GPT-4 offers for agrochemical soil analysis with fertilization recommendations. In this article, we conduct an experiment on the data from the previously mentioned article, but this time using artificial intelligence, namely the GPT-4o model and prompt engineering. The goal is to show the additional possibilities that this combination offers.

2. Previous work

Artificial intelligence has already been used in soil research. One of the researches was done in Indonesia: „The study results indicate that the level of trust of soil science experts in Indonesia to ChatGPT answers has reached above 80%.” [2]. The positive impact of data collection and analysis on harvest growth is presented in [3]. The advantages and disadvantages of chatbots in the field of agriculture are shown in another article: „ChatGPT is a game-changer in the field of agriculture and

allied sectors. Its advanced language capabilities can revolutionize precision farming, crop monitoring, food processing, and livestock management. However, it's crucial to keep in mind that this technology also comes with limitations, such as the need for large data sets, potential biases, and computational resources.” [4].

3. Materials and Methods

3.1. GPT-4o

"GPT-4o (“o” for “omni”) is our most advanced model. It is multimodal (accepting text or image inputs and outputting text), and it has the same high intelligence as GPT-4 Turbo but is much more efficient—it generates text 2x faster and is 50% cheaper." [5]. This model was significant for the authors because it could read data from an image, and the image contains a table with input data. It represents an improvement of GPT-4. In our previous work [1] the input and output was in text format which required us to convert the table data into a text question. After receiving the answer, which was in text format, it was necessary to convert it into tabular format. These conversions were done manually, while GPT-4o allows for a greater degree of automation. This means that the input data can be loaded from the image on which the table is located, and the output can be obtained in textual form specifically in JSON and HTML formats. JSON is suitable for further processing by a machine, and HTML code allows for a human-friendly display.

3.2. Prompt Engineering

In order to maximize the possibilities offered by GPT-4o, we used prompt engineering. The goal was for the model to better understand the intentions we had in our experiment and to obtain optimal outputs in relation to our needs.

"Prompt engineering is the art of crafting effective prompts that guide ChatGPT to generate desired responses. It plays a crucial role in maximizing the effectiveness of the language model by bridging the gap between user intent and model understanding. A well-engineered prompt can significantly improve the quality and relevance of ChatGPT’s output, while a poorly crafted one may lead to unsatisfactory or erroneous responses." [6]

We relied on the source ChatGPT Prompt Engineering for Developers [7] which provides a good overview of prompt engineering principles and tactics. We relied on a tactic known as Ask for a structured output because we wanted to get a text response in HTML and JSON format. This has been achieved.

3.3. Agrochemical soil analysis

Agrochemical soil analysis is a basic prerequisite for successful agricultural production in vegetable, arable and fruit growing [8, 9]. Every year, plants take out a certain type and amount of nutrients from the soil, which must be returned to the soil by fertilizing [10].

Average soil samples consisting of the appropriate number of individual samples (20-25) were taken from plots up to 5 ha. The depth from which to take samples for soil fertility control purposes is 0-30 cm for agricultural and vegetable plant species. Samples were taken with the help of soil sampling hand probes. Agrochemical soil analysis was done in the accredited laboratory of the Agricultural Expert Service (PSS) in Zrenjanin for cadastral plot number 256, cadastral municipality Lukino Selo, area 9652m², pre-crop wheat. Fertilizer recommendations are based upon the previous crop, the crop to be grown and the yield goals (just for N).

4. Results and Discussion

The basis for asking the question is the program code at [11]. Table 4 (the results of soil analysis) is placed inside the image that is placed on the Internet and its address is <https://i.imgur.com/sMnj3aF.png> (Figure 1).

Table 4. Soil analysis results

Parameter	Unit	Value
Soil reaction (pH) in 1M KCl		7,56
Soil reaction (pH) in water		8,46
CaCO ₃ content	%	4,35
Easily accessible phosphorus as P ₂ O ₅	mg/100 gr	27,94
Easily accessible potassium as K ₂ O	mg/100 gr	20,77
Humus content	%	2,59
Total nitrogen content <i>(calculation)*</i>	%	0,10

Figure 1: Image of input data table

Inside the index.js file is the ask method, with two parameters:

1. Text question
2. The image used in the question

Below is a printout of the question asked (user) and the answer received (assistant):

user: Table 4. shows Soil analysis results. We need fertilization recommendations for the following plant species and yield kg/ha: Corn with 10000 kg/ha, Potatoes with 50000 kg/ha, Beans with 2500 kg/ha and Cabbage with 40000 kg/ha. Required are amounts of 3 nutrients: N, P₂O₅ and K₂O, and they should be expressed in kg/ha. Provide them in JSON format with the following keys: plant_specie, yield_kg_per_ha, n, p205, k20. Then transform the obtained results into an HTML table whose columns are: plant_specie, yield_kg_per_ha, n, p205, k20. For image on <https://i.imgur.com/sMnj3aF.png>

assistant: To provide accurate fertilization recommendations, I will need to use standard nutrient uptake figures for each crop. Based on typical requirements, here's the JSON output and corresponding HTML table format for the given plant species and desired yields.

JSON Output

```
```json
[
 {
 "plant_specie": "Corn",
 "yield_kg_per_ha": 10000,
 "n": 250,
 "p205": 100,
 "k20": 250
 },
 {
 "plant_specie": "Potatoes",
 "yield_kg_per_ha": 50000,
 "n": 200,
 "p205": 150,
 "k20": 300
 }
]
```

```

 },
 {
 "plant_specie": "Beans",
 "yield_kg_per_ha": 2500,
 "n": 90,
 "p205": 60,
 "k20": 80
 },
 {
 "plant_specie": "Cabbage",
 "yield_kg_per_ha": 40000,
 "n": 250,
 "p205": 100,
 "k20": 250
 }
]
 ...

```

### HTML Table

```

``html
<table border="1">
 <tr>
 <th>Plant Specie</th>
 <th>Yield (kg/ha)</th>
 <th>N (kg/ha)</th>
 <th>P2O5 (kg/ha)</th>
 <th>K2O (kg/ha)</th>
 </tr>
 <tr>
 <td>Corn</td>
 <td>10000</td>
 <td>250</td>
 <td>100</td>
 <td>250</td>
 </tr>
 <tr>
 <td>Potatoes</td>
 <td>50000</td>
 <td>200</td>
 <td>150</td>
 <td>300</td>
 </tr>
 <tr>
 <td>Beans</td>
 <td>2500</td>
 <td>90</td>
 <td>60</td>
 <td>80</td>
 </tr>
 <tr>
 <td>Cabbage</td>
 <td>40000</td>
 <td>250</td>
 <td>100</td>

```

```

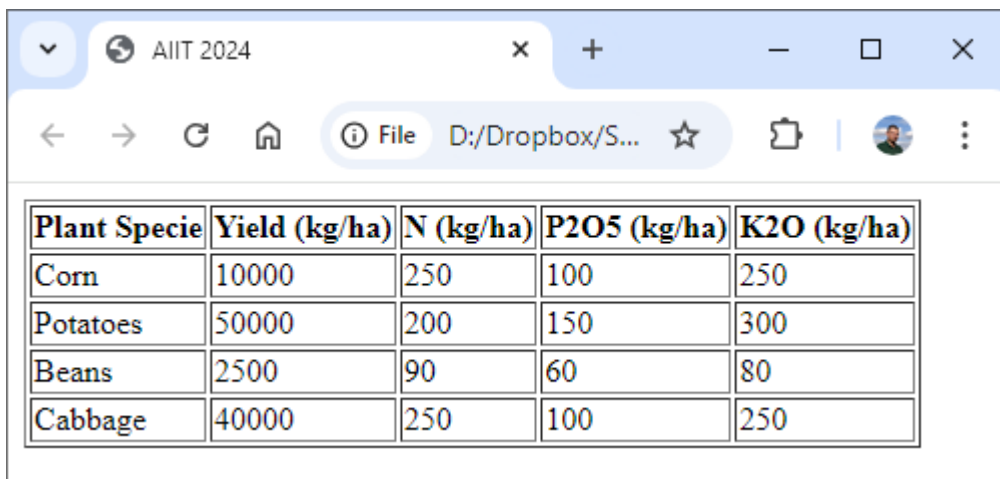
<td>250</td>
</tr>
</table>
...

```

These values are generally accepted and used as a standard for high-yield scenarios. You may need to adjust the values based on specific soil conditions and crop requirements.

Is there any specific data from Table 4. that should affect the recommendation significantly or any additional adjustment factors you would like to apply based on the soil analysis?

When a web page is created from the received HTML code, the following fertilization recommendation result (Figure 2) is obtained on the web page:



Plant Specie	Yield (kg/ha)	N (kg/ha)	P2O5 (kg/ha)	K2O (kg/ha)
Corn	10000	250	100	250
Potatoes	50000	200	150	300
Beans	2500	90	60	80
Cabbage	40000	250	100	250

Figure 2: Result obtained from HTML code fertilization recommendations

The results of soil fertilization recommendation PSS are shown in Table 1. and ChatGPT-4 fertilization recommendations are shown in Table 2.

**Table 1:**

PSS Fertilization recommendation

		Required amount of nutrients by nutritional elements kg/ha		
Plant species	Yield kg/ha	N	P <sub>2</sub> O <sub>5</sub>	K <sub>2</sub> O
Corn	10000	147	79	84
Potatoes	50000	350	216	335
Beans	2500	48	14	21
Cabbage	40000	144	37	112

**Table 2:**

ChatGPT-4 Fertilization recommendation

		Required amount of nutrients by nutritional elements kg/ha		
Plant species	Yield kg/ha	N	P <sub>2</sub> O <sub>5</sub>	K <sub>2</sub> O
Corn	10000	120-150	60-80	80-100
Potatoes	50000	200-250	100-150	200-300
Beans	2500	50-70	40-60	40-60
Cabbage	40000	150-200	80-100	80-100

There was matching of the recommendations for fertilization of PSS and ChatGPT-4 only for the the plant type corn, but it is insufficiently reliable when recommendations for fertilization are considered for vegetable species (potatoes, beans and cabbage). However, the results of HTML code did not provide adequate answers for fertilization recommendations soil.

## 5. Conclusions

We have presented the most important facts related to GPT-4o, prompt engineering and agrochemical soil analysis with fertilization recommendations. Some of the examples of the use of artificial intelligence in the field of agriculture are listed. An experiment was described that had as input a question and a table within an image, based on the capabilities of GPT-4o, and prompt engineering. The experiment gave the result in text form, in two formats: JSON and HTML. The results are discussed at the end and compared with the results from our previous research. Both results were not reliable at the expected level.

Further improvements are possible in terms of using other tactics offered by prompt engineering. There is also the possibility to use some new models that will appear in the future. The field of application does not necessarily have to be only agriculture.

### References:

- [1] M. Drašković, Ž. Eremić, Testing the capabilities of ChatGPT 4.0 for agrochemical soil analysis with fertilization recommendations, *Challenges of Contemporary Higher Education - CCHE*. Vol. 3, pp. 487-492, ISBN 978-86-82744-00-9, Kopaonik, 2024.
- [2] V. Karolinoerita, et al., Application of Chatgpt in Soil Science Research and the Perceptions of Soil Scientists in Indonesia, Electronic copy available at: <https://ssrn.com/abstract=4401008>, 2023.
- [3] E. Elbasi, et al., Crop prediction model using machine learning algorithms. *Applied Sciences*, 13(16), 9288, <https://doi.org/10.3390/app13169288>, 2023
- [4] V. Gaddikeri, M. S. Jatav, J. Rajput, Revolutionizing Agriculture: Unlocking the Potential of ChatGPT in Agriculture, *Food and Scientific Reports*, e-ISSN 2582-5437, Vol. 4, Issue 3, pp. 20-25, 2023.
- [5] GPT-4o, <https://platform.openai.com/docs/models/gpt-4o>, accessed: September 2024.
- [6] S. Ekin, Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. Authorea Preprints, 2023.
- [7] ChatGPT Prompt Engineering for Developers, <https://learn.deeplearning.ai/courses/chatgpt-prompt-eng/lesson/2/guidelines>, accessed: September 2024.
- [8] C.P.Dinkins, C. Jones, Interpretation of soil test results for agriculture, Montana State University Extension, Mont Guide, Publication no. MT200702AG, 2013.
- [9] D. Varga, Priručnik za đubrenje ratarskih i povrtarskih kultura, PSS Subotica AD, 2015.
- [10] E.D. Paolo, M. Rinaldi, Yield response of corn to irrigation and nitrogen fertilization in a Mediterranean environment, *Field Crops Research*, volume 105, pp. 202-210, 2008.
- [11] Gpt-4o-simple-image-question-by-Node.js <https://github.com/ezeljko1981/Gpt-4o-simple-image-question-by-Node.js>, accessed: September 2024.



# A comparative study of Software Development Life Cycle (SDLC) models

Buen Bajrami<sup>1</sup>, Ilija Jolevski<sup>1</sup>, Kostandina Veljanovska<sup>1</sup>

<sup>1</sup> University St. Kliment Ohridski – Bitola, Faculty of Information and Communication Technologies, 1 Maj bb., 7000 Bitola, North Macedonia University, Address, City, Country

bajrami.buen@uklo.edu.mk, ilija.jolevski@uklo.edu.mk, kostandina.veljanovska@uklo.edu.mk

## Abstract:

Software Development life cycle (SDLC) is a framework model used in project management that defines the stages included in an information system and a software development project, from an initial feasibility study to the maintenance of the completed application. There are different software development life cycle models that has to be specified for each software development project. These models are also called "Software Development Process Models". Each process model follows a series of phases unique to its type to ensure success in the steps of its development. We analyze the models one by one, review the latest research and provide comparisons between the Spiral Model, Waterfall Model, Iterative Model, Big Bang Model, V-Model and RAD Model. Then, from the obtained analysis we give an interpretation about their use to customers. And finally, we present the appropriate model for our proposed development project, which according to analysis and research is the most appropriate.

## Keywords:

Spiral Model, Waterfall Model, Iterative Model, Big Bang Model, V-Model, RAD Model

## 1. Introduction

Organizations involved in software development utilize specialized systems that are structured into distinct phases, which outline the processes for developing, maintaining, and improving software. These processes are integral to the SDLC, which plays a crucial role in software project development. Various models exist, each differing in their approach, and organizations select the model that best aligns with their specific requirements. Over the past decades, significant progress has been made in this area, with the development of these models and the continuous emergence of new models that greatly assist engineers in software development. In this paper, we analyze six SDLC models, presenting how they function, their significance, and their practical applications. There are many more SDLC models that can be used for software development, but our research focus only on six of them. We discuss the categories of software development where these models are most applicable and advantageous. Additionally, we explore the future challenges each model may face, highlight areas of strength, and provide recommendations for further improvements.

## 2. Selection of the appropriate model

In order to arrive at the appropriate selection of an SDLC model, we must take into account several factors that will facilitate the comparison and making the decision to continue the web application development process. Therefore, we must base ourselves on some questions that give us a more realistic assessment of the client's requirements and the project as a whole.

- a) How realistic and sustainable are the customer's requirements?
- b) What is the size of the project?
- c) What are the access levels of the project?
- d) What is the dynamic development plan of the project?

The purpose of this paper is to build a scenario taking into account the client's requirements, about the development of a web-based application. The client wants a web application which will function as an online shop, where people will be able to order food online in a restaurant. Also, the customer requests that it be possible for a restaurant customer to make a reservation for a table, at most two weeks in advance. Where all orders and reservations will be stored in the web application database. Below we have described the functions of our project through a use case diagram.

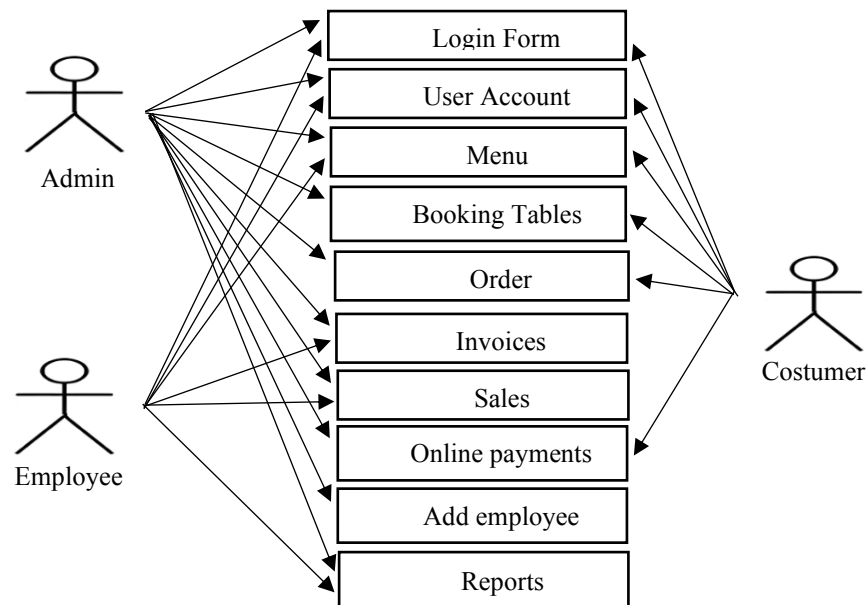


Figure 1. Restaurant online platform - Use Case Diagram

In this paper, we have taken as a basis a total of six models of software development. They are: Spiral Model, Waterfall Model, Iterative Model, Big Bang Model, V-Model, RAD Model.

### 3. Literature review

In the literature review, we have analyzed all the models included in the research, presenting the orientation of the development and use of these models. Software engineering involves the analysis and design of a software system that handles one or several specific tasks. According to [1] analysis describes the “what” of a software system, which means what happens in the current and what will be required in the new software system. This refers to requirement analysis or gathering. In recent time, the most popular methodological approaches for developing software for a computer-based information system are the popular traditional Waterfall Model [2]. For example [3] the author created a simulation system of the Waterfall model, which was able to help project managers in determining how to achieve maximum productivity with the minimum number of hours, costs and workers. The authors [4] have conducted a comparative study between the iterative waterfall life cycle model and the incremental software development cycle model for resource optimization using the Symphony.NET simulation tool. Spiral model [5] is similar to the incremental model, with more emphasis placed on risk analysis. As for the Big Bang model, the authors [6] show that the Big Bang Model is one in which a large amount of people or money is gathered, a lot of energy is expended, and the perfect software product comes out or not. No matter which model we choose there is possibility to include generative artificial intelligence (GenAI) which can bring unparalleled enhancements in various phases of SDLC. Enterprises could embrace GenAI in their software development process in order to stay competitive in the huge software market. By integrating Gen AI quicker development cycles, superior code quality, and increased innovation could be realized [7].

## 4. Comparative study of SDLC models

In this paper, we research each model and addressed them through three main factors. The first factor examines the general data of a model, including details of its functioning and the reasons for selecting a particular model. The second factor focuses on the application of an SDLC model in different areas of software engineering. Finally, the third factor addresses the challenges of the models by analyzing their respective strengths and weaknesses.

### 4.1. Spiral Model

The Spiral model is a software development life cycle (SDLC) model used for risk management, combining elements of the iterative and Waterfall models. This model is particularly suitable for large, complex, and extensive projects. A notable feature of this model is that it releases a prototype after each phase of the spiral, which is then refined. It is effective in managing risks by continuously creating and testing prototypes, allowing for an analysis of their strengths and weaknesses. Below is an illustration depicting the main stages of the Spiral model [8].

The Iterative model consists of four stages: planning of objectives, risk analysis, engineering or development, and review. A project repeatedly passes through all these stages, with the phases forming a spiral within the model [9].

1. Determine objectives and find alternate solutions. This phase includes requirement gathering and analysis. Based on the requirements, objectives are defined and different alternate solutions are proposed.
2. Risk Analysis and resolving – In this quadrant, all the proposed solutions are analyzed and any potential risk is identified, analyzed, and resolved.
3. Develop and test: This phase includes the actual implementation of the different features. All the implemented features are then verified with thorough testing.
4. Review and planning of the next phase – In this phase, the software is evaluated by the customer.

#### 4.1.1. Spiral Model Risk Analysis and Model Application

The Spiral model analyzes the proposed results and systematically identifies, analyzes, and addresses all possible risks. Following this, methods such as prototyping, simulation, benchmark testing, analytical models, and user research are employed to develop the lowest-risk and most cost-effective strategy [10]. At this stage of the project, the product manager refers to the risk register to identify all potential risks associated with the spiral's objectives and requirements. These risks may be technical, financial, market-related, operational, or environmental. Consequently, risk registers are created with the aim of investigating or addressing these risks in both the short and long term. A risk register is an essential component of various product management documents and templates, including discovery documents and written epics [11].

The spiral model is one of the most widely recognized Software Development Life Cycle (SDLC) models and is applied in a broad range of projects. This model is particularly well-suited for projects that require frequent releases and those where changes may be requested at any stage of the process. Additionally, it is ideal for projects with medium to high risk levels and for those where cost and risk analysis plays a critical role. The spiral model is also highly beneficial in projects with unclear or complex requirements, where a more flexible and adaptive approach is necessary to address uncertainties and challenges effectively.

### 4.2. Waterfall Model

The Waterfall model is a classical approach used in the system development life cycle, characterized by a linear and sequential process. It is termed "Waterfall" because the model progresses systematically from one phase to the next in a downward manner. This model is divided into distinct phases, where

the output of one phase serves as the input for the next. Each phase must be completed before the subsequent phase begins, with no overlap between phases [12].

All the stages of project development following the model Waterfall are [13]:

1. Requirement Gathering and analysis – All possible requirements of the system to be developed are captured in this phase and documented in a requirement specification document.
2. System Design – The requirement specifications from first phase are studied in this phase and the system design is prepared.
3. Implementation – With inputs from the system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality, which is referred to as Unit Testing.
4. Integration and Testing – All the units developed in the implementation phase are integrated into a system after testing of each unit. Post integration the entire system is tested for any faults and failures.
5. Deployment of system – Once the functional and non-functional testing is done; the product is deployed in the customer environment or released into the market.
6. Maintenance – There are some issues which come up in the client environment. To fix those issues, patches are released.

#### **4.2.1. Waterfall Model Risk Analysis and Model Application**

The risks associated with the Waterfall methodology can be mitigated or eliminated through effective risk management. In Waterfall, risks are managed using charts that track the risk from the moment it is detected. Managers monitor its progression, and if the risk intensifies as the project advances, its line on the chart rises. In such cases, developers and managers work to identify and replace the product features contributing to the risk. Conversely, if the risk line decreases as the project progresses, the risk is considered to have low priority. Risk charts in Waterfall are divided into three sections: the first begins when the risk is detected and ends when developers start seeking a solution; the second phase continues until a solution is found; and the third phase, the risk removal phase, concludes on the risk's expiry date. [14].

Every software project is unique and requires a tailored Software Development Life Cycle (SDLC) approach based on various internal and external factors. The Waterfall model is particularly suitable in certain situations. It is most effective when the project requirements are well-documented, clear, and fixed from the outset. The model is also ideal when the product definition remains stable throughout the development process and when there are no ambiguous or unclear requirements. Furthermore, the Waterfall model works well when there are sufficient resources with the necessary expertise to support the development. Finally, it is best suited for short-term projects where a structured, linear approach is more efficient.

#### **4.3. Iterative Model**

This model does not start with a complete set of requirements. Instead, the design and development process begins with the basic requirements. Once the product is developed according to these initial requirements, it undergoes a review to identify or define additional requirements. Upon completion of the review phase, the current iteration concludes. Based on this review, the next set of requirements is determined, and the design and development of the second iteration begin. Once the product is developed according to the requirements of the second iteration, it is reviewed again to identify or define further requirements. These iterations continue until the final product is developed [15].

1. Requirement Gathering and Analysis - Business requirements are gathered during this phase of the Iterative model. Then, an analyst determines whether they can be met within financial constraints.
2. Design - During this phase of the iterative model, the project team receives the complete list of criteria to begin work in a specific direction. They then utilize various diagrams, such as data flow

diagrams, class diagrams, activity diagrams, state transition diagrams, and others, to gain a clear understanding of the program design and to assist them in progressing with development.

3. Implementation - At this point in the project, according to the iterative model, the actual coding of the system begins. This phase is influenced by the analysis and design from the Design phase. All requirements, plans, and design strategies have been executed.

4. Testing - This phase involves comparing the current build iteration against a set of rules and standards to determine whether it meets them. This type of testing includes performance testing, stress testing, security testing, requirements testing, usability testing, multi-site testing, disaster recovery testing, and others.

5. Deployment - After all phases are completed, the software is deployed to the working environment.

6. Review - In this phase, after the product has been deployed, the behavior and validity of the deployed product are checked. If any errors are found, the process begins again from the requirement gathering stage.

7. Maintenance - In the maintenance phase, after the software has been deployed in the working environment, bug fixes or new updates may be required.

#### 4.3.1. Iterative Model Risk Analysis and Model Application

In the Iterative model, risk management is not just a component of project management; it is the essence of project management. It focuses on minimizing negative outcomes while balancing the achievement of desirable ones: preventing certain adverse events while also ensuring that the project's positive risks are realized [16]. The process of managing risk in the Iterative Model involves several key steps to ensure that risks are identified, analyzed, and controlled effectively throughout the development lifecycle. The first step is risk management planning, where we determine who needs to be involved and establish the timing and frequency of risk management activities. Next, during risk identification, we analyze potential risks and opportunities across various business categories. In the quantitative risk analysis phase, we subjectively assess the probability and impact of these risks, identifying those that are critical. Then, in the qualitative risk analysis stage, we prioritize which areas require time and resources for risk management. Following this, risk response planning is conducted to determine the best strategies to reduce the likelihood and impact of identified risks. Finally, in risk monitoring and control, the risk response plan is implemented, with ongoing management to ensure progress and compliance.

The Iterative model finds several productive applications across various fields, where its flexibility and adaptability offer significant advantages. In digital marketing, for instance, the model is used to create more engaging advertisements. By conducting thorough analysis to understand user requirements, iterative cycles allow for the continuous improvement of advertisement designs. The model is also applicable in the development of advanced technologies, such as smartphones, cars, and electronic devices, where ongoing iterations enable gradual enhancements and updates. Furthermore, engineering teams often rely on the Iterative model when developing new features or implementing problem-fixing techniques [2]. These iterations, which may not be visible to end users, allow teams to refine their products internally before release, ensuring higher quality and functionality.

#### 4.4. BIG BANG model

In the Big Bang model, the developers of this model do not follow any specific process. Only the necessary funds are enough to start the project. The result of this model may not meet the client's requirements because the client's requirements are not concrete and complete either. This model is suitable for small projects where a very small number of developers can work. [17]

The Big Bang model in software development consists of three key phases [17]:

1. Minimal Planning: Little to no planning is done. Developers begin coding with only a general idea of the final goal, without detailed requirements.

2. Chaotic Development: Components are developed and integrated in an ad-hoc manner, with minimal testing and uncontrolled changes during the process.
3. Delivery or Reorganization: The product is either completed or delivered or fails to meet requirements, leading to a full reorganization or restart of the project.

#### 4.4.1. BIG BANG Model Risk analysis and Model Application

The Big Bang model is a software development life cycle (SDLC) approach where the development process begins with little or no planning and all the modules or components are developed simultaneously. Risk analysis in the Big Bang model involves identifying potential risks and assessing their impact on the project. The Big Bang model, while simple in its approach, presents several risks that must be carefully analyzed to avoid potential pitfalls [24]. One major concern is the lack of requirements definition, where unclear or insufficient requirements can lead to misunderstandings, scope creep, and difficulties in delivering a product that meets user expectations. Another challenge is integration, as in the Big Bang model, modules or components are developed independently and only integrated at a later stage, which can lead to unexpected issues. Additionally, insufficient testing coverage is a common risk, as the simultaneous development approach may overlook thorough testing. This also increases the effort required for debugging, as the simultaneous development of modules raises the likelihood of defects and bugs. Moreover, the model offers limited scope for change control, making it difficult to accommodate changes to requirements or design once development is in progress. Lastly, scalability and performance issues can arise, as these considerations may not be adequately addressed during the initial stages of development in the Big Bang model.

In this model, minimal time is spent on detailed planning, and development begins directly with the required funds and efforts as inputs. This approach is well-suited for short-term projects, such as academic or practical endeavors. It is particularly effective for small teams, usually consisting of two or three developers, working collaboratively. However, this paradigm is not suitable for large-scale or complex software development due to its high risk, though it may be used for temporary, experimental, or very small software projects. It is typically ideal for smaller projects with small development teams focusing on constructing a simple application or prototype. Additionally, the model works well for academic, learning, or practice projects, offering a flexible and low-cost framework. Finally, it can be useful for software products with unclear or poorly understood requirements, particularly when no specific release date is established, allowing for ongoing experimentation and development flexibility.

#### 4.5. V-Model

V-model means Verification and Validation model. Just like the waterfall model, the V-Shaped life cycle is a sequential path of execution of processes. Each phase must be completed before the next phase begins. V-Model is one of the many software development models.

The V Model contains Verification phases on one side and Validation phases on the other. The implementation/coding phase joins the verification and validation phases in V-shape. Thus it is called a V Model. We present below both phases: [18]

##### **Verification phases:**

1. Business requirement analysis: This is the first step where product requirements understood from the customer's side. This phase contains detailed communication to understand customer's expectations and exact requirements.
2. System Design: In this stage system engineers analyze and interpret the business of the proposed system by studying the user requirements document.
3. Architecture Design: The baseline in selecting the architecture is that it should understand all which typically consists of the list of modules, brief functionality of each module, their interface relationships, dependencies, database tables, architecture diagrams, technology detail, etc.
4. Module Design: In the module design phase, the system breaks down into small modules. The detailed design of the modules is specified, which is known as Low-Level Design

5. Coding Phase: After designing, the coding phase is started. Based on the requirements, a suitable programming language is decided. There are some guidelines and standards for coding.

#### **Validation phases:**

1. Unit Testing: In the V-Model, Unit Test Plans (UTPs) are developed during the module design phase. These UTPs are executed to eliminate errors at code level or unit level.
2. Integration Testing: Integration Test Plans are developed during the Architectural Design Phase.
3. System Testing: System Tests Plans are developed during System Design Phase.
4. Acceptance Testing: Acceptance testing is related to the business requirement analysis part. It includes testing the software product in user atmosphere. Acceptance tests reveal the compatibility problems with the different systems, which is available within the user atmosphere [18].

### **4.5.1. V-Model Risk Analysis and Model Application**

The V-Model is a highly structured approach to software development that involves a series of phases, each of which is associated with a specific set of deliverables. The V-Model is often used in safety-critical industries, such as aerospace and defense, where the consequences of software failure can be catastrophic. Risk management is an essential component of the V-Model methodology. It involves identifying potential risks and taking steps to mitigate or eliminate them. The V-Model provides a framework for risk management by breaking the development process down into a series of phases, each of which is associated with a specific set of deliverables. It provides a structured approach to identifying and mitigating potential risks, which can help to ensure that the project is completed on time, within budget, and to the required quality standards [19].

It is advisable to use the V Model on small-to-medium-sized software projects where the requirements are clear without any ambiguity. For projects where acceptance criteria are proper, V Model is the preferable choice. The V Model is useful when ample technical resources are available with technical expertise, and tech stacks and tools are not dynamic [20].

### **4.6. RAD Model**

Rapid Application Development or RAD means an adaptive software development model based on prototyping and quick feedback with less emphasis on specific planning. In general, the RAD approach prioritizes development and building a prototype, rather than planning. With rapid application development, developers can quickly make multiple iterations and updates to the software without starting from scratch. This helps ensure that the final outcome is more quality-focused and aligns with the end users' requirements [22].

Below we have presented the development stages of the RAD model [22]:

1. Business Modeling: On basis of the flow of information and distribution between various business channels, the product is designed
2. Data Modeling: The information collected from business modeling is refined into a set of data objects that are significant for the business
3. Process Modeling: The data object that is declared in the data modeling phase is transformed to achieve the information flow necessary to implement a business function
4. Application Generation: Automated tools are used for the construction of the software, to convert process and data models into prototypes
5. Testing and Turnover: As prototypes are individually tested during every iteration, the overall testing time is reduced in RAD.

#### 4.6.1. RAD Model Risk Analysis and Model Application

Risk analysis in the Rapid Application Development (RAD) Model of the Software Development Life Cycle (SDLC) focuses on identifying potential risks and assessing their impact on the project's success. One of the key risks is inadequate requirements gathering, as the fast-paced nature of RAD can lead to incomplete or inaccurate understanding of the project's needs. Additionally, resource availability is critical, as this model demands a dedicated and highly skilled team to meet the rapid development timelines. Technical complexity is another concern, especially in projects that involve complex integrations or advanced technical components. Customer involvement is essential in this model, as continuous feedback and active decision-making are crucial for success. A lack of customer participation can lead to misunderstandings, delays, and misaligned project goals. Integration and compatibility issues often arise in RAD projects, especially when integrating new systems or components with existing infrastructure. Furthermore, testing and quality assurance can be compromised due to the limited time available for comprehensive testing, increasing the risk of defects or performance issues. Lastly, security and data privacy are potential risks, as rapid development may overlook critical security measures, leading to vulnerabilities and the possibility of data breaches. These risks must be carefully managed to ensure the success of a RAD-based project [4, 22].

When deciding to use the Rapid Application Development (RAD) model, it is important to evaluate the specific project needs, as this model may not be productive for unsuitable demands. This model is best applied in cases where the system requirements are well-defined and a short development timeline is required. It is also effective for projects that can be modularized, with components that are reusable or readily available for development. Additionally, the model is suitable when existing system components can be leveraged to build a new system with minimal changes. However, this model should only be used when teams consist of domain experts, as extensive knowledge and the ability to apply advanced techniques are crucial for successful implementation. Finally, RAD is appropriate when the project budget allows for the use of automated tools and specialized techniques necessary for the rapid development process.

### 5. Discussion

Taking into account the model characteristics, risk analysis, possibilities of application we can summarize what are the challenges for each model.

#### *Spiral model challenges*

The spiral model presents several challenges that need to be considered before its implementation. One major drawback is the high cost, making it unsuitable for small projects due to the extensive resources required. Additionally, the model relies heavily on risk analysis, necessitating the involvement of experts in risk management at every stage of the project. Its complexity is another significant challenge [5], as it is the most intricate SDLC model, involving extensive documentation and multiple intermediate phases. Furthermore, managing time can be difficult, as the number of phases is not predetermined at the start of the project. This uncertainty complicates time planning and may result in an overall increase in the project budget.

#### *Waterfall model challenges*

The Waterfall model is one of the earliest software development models, but it comes with several challenges that should be carefully evaluated when selecting the appropriate approach for a project. One of the main drawbacks is that no working software is produced until the later stages of the development lifecycle. This leads to high levels of risk and uncertainty, making it unsuitable for complex or object-oriented projects. Additionally, it is a poor fit for long or ongoing projects, as it does not accommodate changes in requirements well, particularly when there is a moderate to high risk of such changes [12]. Measuring progress during the various stages can be difficult, as each phase must be completed before moving to the next. Furthermore, integration is performed as a "big-bang" at the end of the project, which delays the identification of potential technological or business challenges until the final stages, limiting the ability to address them early in the process.



***Iterative model challenges***

The Iterative model, while highly flexible and adaptive, comes with several challenges that need to be considered. First, it may not be suitable for smaller projects, as the overhead associated with iterative cycles can outweigh the potential benefits. Additionally, more resources in terms of time and money may be required due to the repeated cycles of development, testing, and feedback, making it resource-intensive. Changes in requirements during the development process can also lead to budget overruns, as accommodating these changes often requires additional time and effort. Furthermore, the iterative approach may not be ideal for projects with strict or inflexible requirements, as the model's continuous changes and improvements may not be feasible in such environments. Lastly, effective communication between the development team and users is crucial for the success of the Iterative model, as ongoing feedback and collaboration are essential to its process.

***BIG BANG model challenges***

The Big Bang model faces several future challenges that need to be addressed to enhance its effectiveness. First, the client's requirements must be very clear from the outset, as the lack of thorough planning can lead to misunderstandings and deviations from the desired outcome. Second, it is not suitable for large projects, where the complexity and scope exceed the capabilities of this simplistic approach. Lastly, the model does not offer high security, making it a risky choice for projects where data protection and security are critical. These limitations highlight the need for further refinement and adaptation to make the model more robust and reliable in a broader range of applications.

***V-Model challenges***

Similar to other models discussed, the V model also presents several challenges that should be carefully considered when deciding to use it for project development. One significant drawback is that the model is very rigid and offers minimal flexibility, which can be problematic for projects that require adaptability. Additionally, the V model is not ideal for complex projects, as its linear structure may not accommodate the dynamic needs of more intricate systems. Another challenge is that software is only developed during the implementation stage, meaning there are no early prototypes available for review or testing. Lastly, if changes occur mid-project, both the test and requirement documents must be updated, adding to the complexity and potential delays in the project timeline. These challenges highlight the importance of thorough analysis before choosing the V model for a development project.

***RAD Model challenges***

Like all SDLC models, the RAD model has its own set of challenges that must be considered before choosing it for a project. One significant challenge is that powerful and efficient tools used in RAD require highly skilled professionals, making it essential to have a team with the right expertise. Another issue is that the absence of reusable components can jeopardize the project's success, as RAD relies heavily on reusing existing elements. Additionally, the team leader must collaborate closely with both developers and customers to ensure the project is completed on time. This model is unsuitable for systems that cannot be modularized, as the model depends on breaking down the system into modules. Moreover, customer involvement is required throughout the entire development life cycle, which can be challenging to maintain. Lastly, RAD is not appropriate for small-scale projects, as the costs of automated tools and advanced techniques may exceed the overall project budget, making it impractical for smaller projects.

***Discussion of models comparison***

The SDLC models comparison in this paper was done for a small-scale project that does not require significant costs or a large number of developers. Other models are better suited for larger projects, which demand higher costs, more developers, additional tests or prototypes, and more time. Consequently, we believe that the Big Bang model is the most suitable for developing our web application. This model requires a small team of developers and a relatively short time frame for project completion, provided that our requirements for the web application are clear and precise, leaving no room for confusion or challenges during the process. Therefore, we recommend using the Big Bang model for small projects, with a focus on improving security as a future challenge. This will enhance the project's practicality and seriousness, ensuring no deficiencies in any aspect. Below, we present all the SDLC models based on Table 1, where we evaluated and compared each model using six performance factors.

Table.1.The comparison of SDLC Models

Features	Spiral Model	Waterfall Model	Iterative Model	Big Bang Model	V-Model	RAD Model
Understanding Requirements	Medium	High	Medium	Medium	High	High
Cost	High	Low	High	Low	High	High
Risk Involvement	High	High	Low	High	High	High
User Involvement	Medium	Low	High	Medium	High	Medium
Guaranty of Success	High	Low	Low	Low	Medium	High
Flexibility	High	Low	Medium	Medium	Low	High

## 6. Conclusion

In conclusion, Software Development Life Cycle (SDLC) models serve as essential frameworks that guide the development process of software applications. After conducting a comprehensive review of the literature and analyzing each model individually, we have been able to assess which model is most suitable for web application development. While each SDLC model may be appropriate and functional for specific projects, the primary objective of this research was to identify the most appropriate model for our project by comparing four key parameters across each model. These parameters have enabled us to determine which model offers the optimal solution for our specific case. It is imperative that the choice of an SDLC model is carefully aligned with project goals, stakeholder needs, and the dynamic nature of the software development landscape. As technology continues to evolve, the adaptability and flexibility of SDLC models remain critical for ensuring the success of software projects in meeting user expectations and achieving business objectives. By engaging in continuous evaluation, refinement, and the adoption of best practices, organizations can effectively leverage SDLC models to innovate, collaborate, and deliver value in an ever-changing digital environment. Nevertheless, it is crucial to consider the limitations and challenges inherent in each model and to focus on their ongoing improvement and development. Additionally, developers should explore the potential integration of generative artificial intelligence within the SDLC. This inclusion could significantly transform the SDLC process, particularly in how software is planned, developed, delivered, and maintained. Generative AI holds the potential to automate repetitive tasks performed by developers, testers, and machines, thereby streamlining code generation and other related processes.

### References:

- [1] Bennett, S., McRobb, S. & Farmer, R., Object- Oriented Systems Analysis and Design Using UML, McGraw-Hill Higher Education, 2002.
- [2] Pfleeger, S.L. et al, Software Engineering: Theory and Practice, 3rd Ed. US:Prentice Hall., 2006.
- [3] Y. Bassil, ""A Simulation Model for the Waterfall Software Development Life cycle"," International Journal of Engineering&Technology, pp. Vol.2, no.5, pp. 2049-3444, 2012.
- [4] Prakriti Trivedi et al., ""A Comparative Study between Iterative Waterfall and Incremental Software Development Life Cycle Model for Optimizing the Resources using Computer Simulation"," Information Management in the Knowledge Economy (IMKE), pp. Vol. 7,No.5,pp. 188-194. , 2013.
- [5] B. B, ""A Spiral Model of Software Development and Enhancement"," ACM SIGSOFT Software Engineering Notes", "ACM", 1986.
- [6] Wollack, Edward J., ""Cosmology: The Study of the Universe"," Universe 101: Big Bang Theory. NASA, 2010.

- [7] H. Padmanaban, et al., "Implication of Artificial Intelligence in Software Development Life Cycle: A state of the art review", International Journal of Recent Research Aspects, vol. Vol. 6, no. Issue 2, pp. 93-928, 2019.
- [8] TechTarget, 2019. URL: [www.techtarget.com/searchsoftwarequality/definition/spiral-model](http://www.techtarget.com/searchsoftwarequality/definition/spiral-model).
- [9] Artoftesting, "www.artoftesting.com," 2023. URL: <https://artoftesting.com/spiral-model>.
- [10] Shiksha, "www.shiksha.com," 2023. [Online]. Available: <https://www.shiksha.com/online-courses/articles/spiral-model-in-software-engineering/>.
- [11] Logrocket, "www.blog.logrocket.com," 2023.: <https://blog.logrocket.com/product-management/how-to-create-an-effective-risk-register-like-a-product-manager/>.
- [12] Economictimes, 2024. URL: <https://economictimes.indiatimes.com/definition/waterfall-model>.
- [13] Tutorialspoint, 2023. URL: [https://www.tutorialspoint.com/sdlc/sdlc\\_waterfall\\_model.htm](https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm).
- [14] Hygger, "www.hygger.io," 2023.: <https://hygger.io/blog/the-risks-of-waterfall-methodology/>.
- [15] Tc, "www.tutorialscampus.com," 2023. [Online]. Available: <https://www.tutorialscampus.com/sdlc/iterative-model.htm>.
- [16] Netlogx, "www.netlogx.com," 2023. [Online]. Available: <https://netlogx.com/blog/2022/05/17/iterative-risk-management-methodology/>.
- [17] Javatpoint, "www.javatpoint.com," 2023. [Online]. Available: <https://www.javatpoint.com/software-engineering-big-bang-model>.
- [18] Javatpoint, 2024. [Online]. Available: [www.javatpoint.com/software-engineering-v-model](http://www.javatpoint.com/software-engineering-v-model).
- [19] Medium, 2024.: [medium.com/smart-project-kit/risk-management-in-v-model-methodology-62c65ab2154b#:~:text=The%20V%2DModel%20and%20Risk%20Management&text=V%2DModel%20methodology.,It%20involves%20identifying%20potential%20risks%20and%20taking%20steps%20to%20mitigat](https://medium.com/smart-project-kit/risk-management-in-v-model-methodology-62c65ab2154b#:~:text=The%20V%2DModel%20and%20Risk%20Management&text=V%2DModel%20methodology.,It%20involves%20identifying%20potential%20risks%20and%20taking%20steps%20to%20mitigat)
- [20] Testsigma, 2024. URL: <https://testsigma.com/blog/v-model-in-software-development-life-cycle/>.
- [21] Javatpoint, 2024. URL: <https://www.javatpoint.com/software-engineering-incremental-model>.
- [22] Kissflow, 2024. URL: <https://kissflow.com/application-development/rad/rapid-application-development/>.
- [23] tutorialspoint, 2023. URL: [https://www.tutorialspoint.com/sdlc/sdlc\\_waterfall\\_model.htm](https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm).
- [24] geeksforgeeks, 2024. URL: <https://www.geeksforgeeks.org/big-bang-integration-testing/>.

# Program Model for a Visual Editor of Gamepad Haptic Effects

Vasilije Bursać<sup>1</sup>, Dragan Ivetić<sup>1</sup> and Aleksandar Kupusinac<sup>1</sup>

<sup>1</sup> *University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia*

*vasilijebursac@uns.ac.rs; ivetic@uns.ac.rs; sasak@uns.ac.rs*

## Abstract:

Haptic feedback improves immersion of video games by using vibration to stimulate the player's sense of touch and thus enhance his experience of the events that occur in the game world. The experience of using game engines in teaching across several university courses over the past 14 years has indicated the need for a more efficient approach to designing and testing haptic effects. Most popular game engines provide only basic built-in functions for creating gamepad vibrations. More advanced functionalities are available through additional paid third-party assets or are exclusively accessible to licensed developers during the development of games for specific consoles. A generic haptic feedback model would enable game creators to focus on designing realistic haptic effects instead of technical implementation. This paper presents one program model implemented in the Unity engine, functional on all gamepads and platforms, easily integrated into games, providing a multimodal approach to developing reusable haptic effects: through a user-friendly graphical interface within the engine, from audio files, or programmatically. The described model can be extended to support mobile phone vibration and integration with other feedback systems. Finally, its underlying philosophy can be transferred to other game engines and other devices capable of providing haptic feedback.

## Keywords:

gamepad vibration/rumble, haptic feedback, game development, immersion, Unity

## 1. Introduction

The immersiveness of the video game is its ability to “draw” the player into its virtual world and create a strong sense of presence and engagement by allowing the player to experience the game as if they were “in the skin” of the main character. Therefore, it is probably the most desirable feature of a video game that largely guarantees its success and good reactions from players. In immersive video games, the feedback provided by the game through monitor, speakers, and controller replaces the player's real senses, creating virtual senses that enable the player to experience and feel events and objects in the game as if they were in his physical reality [1].

Over the years, the visual and audio aspects of video games have reached a remarkably high level of fidelity, practically becoming an extension of the player's senses of sight and hearing in the video game world. The quality of graphics and visual effects is constantly becoming more impressive and realistic, and the abundance of quality audio effects has become the standard. In order to push the limits of immersion and provide players with an even more realistic, tangible, and intimate experience, video games rely on haptic feedback to stimulate the sense of touch.

In the context of application in video games, haptic feedback is provided through different types of video game controllers that stimulate the muscles, joints, and receptors in the user's skin using a wide range of stylized vibrations. Their purpose is to amplify the simulation of the user's interaction with objects and events from the virtual world of the video game, with the aim of making the player feel more connected to the game. Studies have shown [2] that haptic feedback directly affects the immersiveness of a video game, which is especially manifested through the player's awareness of his surroundings in the game. For instance, haptic feedback can be utilized to simulate different forces that directly or indirectly affect the player, to emphasize dramatic situations, or to direct the player's attention to specific events and actions of non-playable characters in the game [3].

The most common and cost-effective method of providing haptic feedback is through the vibration of the gamepad controller, i.e. gamepad rumble, which is generated by several vibration motors built into the gamepad that can operate at varying speeds. Although most gamepads nowadays contain some type of vibration motors, not all games support haptic feedback. Video games must contain explicitly implemented programming functions to provide haptic feedback, as well as various haptic effects, i.e. rumble patterns that will be played to the player when certain actions in the game occur. In order to ensure the correct use of haptic feedback and achievement of its benefits, it is necessary to carefully design and synchronize the haptic effects with events in the video game. Additionally, new functionalities and contents that are subsequently added to the game sometimes require the creation of new adequate haptic effects. The development of such a system requires additional time investment by the development team, which can often be challenging to fit into the deadline or budget for video game development. As a result, many games do not support haptic feedback at all or only feature a small number of basic haptic effects.

Research indicates that a prototyping tool with a graphical user interface greatly facilitates the design and testing of haptic vibrotactile effects, particularly when compared to creating these effects directly through writing code [4]. Similar findings emerged from our teaching in several university courses over the past 14 years, including Game Development Process, Virtual Reality Systems, Visual Programming for Animation, and Scripting Languages for Computer Animation. In these courses, students develop video games, simulations, and various different virtual experiences using popular game engines. Work on lab exercises indicated that the current approach to designing and testing haptic effects in popular game engines could be improved to enhance efficiency, intuitiveness, and overall effectiveness, particularly in a university environment where time for lab exercises is limited.

This paper aims to present a flexible program model for the rapid implementation of gamepad vibration in games created using the Unity engine. The developed model enables fast and efficient prototyping, creation and adjustment of haptic effects through the streamlined user interface directly inside of the game engine, without the need for any additional software. One distinctive advantage of the solution presented in this paper is its flexibility, because it provides a multimodal approach to haptic effect creation. Specifically, users can manually create haptic effects through the graphical user interface, generate them from audio files, or programmatically in real time using predefined functions. These methods for effect creation can be combined depending on the use cases, requirements, and user preferences. Created effects are stored as assets within the video game project and they can be copied and reused in other video games.

The paper is divided into five sections. The second section reflects on related work in this area and outlines the strengths and weaknesses of existing program support for gamepad vibration in the most popular game engines. Design and implementation of the program model are described in detail in section three, while section four identifies its limitations and suggests areas for future improvements. The fifth section concludes the paper.

## 2. Previous work

Before the implementation of the proposed solution, an analysis was conducted on existing solutions in the most popular game engines, their assets and packages, external software tools, and scientific papers. The findings of this analysis are presented in the rest of this section. These conclusions were used to determine the fundamental requirements and desirable features for the proposed solution.

In general, most popular game engines provide only basic built-in functions for creating gamepad vibrations, while more advanced functions can be eventually added through assets and packages available for the specific engine. Additionally, sometimes more advanced features exist within the engine, but are platform dependent and exclusively available to licensed developers during the development of games for specific consoles and their controllers. In their native form, without use of any additional assets, Unity and Godot offer only the most basic functions for vibration, which can be used as a basis for the implementation of more advanced and comprehensive vibration algorithms.

Unity engine provides basic functions for controlling gamepad vibration, including adjusting the speeds of the left and right motors, pausing, resuming, and completely stopping the vibration [5]. Next, several packages for the Unity engine that extend the range of its functionalities for gamepad vibration

will be described. These assets rely on Unity's new Input System package. It is important to note that these assets are not free and are licensed on a 'per-seat' basis, meaning that each asset can be installed on only one computer. A separate license must be purchased for each computer, employee, or student that will use it.

Haptic package by Abthoclo [6] for Unity enables easy integration and design of haptic feedback in PC and console games. It features a simple user interface for creating haptic effects and a visual preview within the Unity editor (Figure 1a). Users can design haptic effects by adjusting duration and intensity, i.e. gamepad motors' speeds using curves. The preview section shows curves for the left and right vibration motors, plays both curves simultaneously on the gamepad, and includes a playhead with start and stop options for controlling playback of the preview.

Nice Vibrations package for Unity is a powerful solution that offers a lot of features for adding haptic feedback to iOS, Android, PC, and console games. The package contains a set of sample haptic clips and includes functions for controlling playback and adjusting the intensity or sharpness of effects at runtime. It allows playing native haptic presets, creating transient and continuous haptic effects by specifying intensity and sharpness, and using advanced haptic patterns defined by AHAP (Apple Haptic and Audio Pattern) files or waveform assets [7]. For continuous haptics, intensity and sharpness can be specified with curves. Advanced haptic patterns intended for gamepads are saved as RumbleWaveForm assets. These assets have a precise, but complex and cumbersome format (Figure 1b), making manual editing tedious for users. Therefore, the most convenient approach to creating advanced patterns for gamepad vibration is by converting AHAP files. Designing AHAP files or generating them from audio clips can be done using external software tools like Haptrix [8], Meta Haptic Studio [9], and others. However, if users need to adjust the gamepad haptic effect after generating the RumbleWaveForm asset from an AHAP file, they often need to repeat the entire process due to the inconvenient format of RumbleWaveForm assets. Additionally, previewing haptic effects within the editor is not possible without running the project. There is no visual representation of the effect with a playhead to indicate which part of the effect is currently playing, making it challenging to identify where effect needs to be adjusted. Consequently, users must rely on external software tools to design advanced haptic effects or adjust them after they have been generated from audio and AHAP files.

Godot provides similar functions as Unity, with the exception that it does not include separate functions for pausing and resuming the vibration. However, during the initialization of the vibration, in addition to motor speeds, the total duration of the vibration can be optionally declared [10]. There are still no packages and assets that extend the engine functionality related to haptic feedback and controller vibration. This is primarily due to the fact that Godot is still relatively new, its community, user base, and market are smaller, and the engine is more frequently used among hobbyists and indie developers than in the professional industry.

Unreal Engine provides the most complete and advanced built-in system for implementing gamepad vibration among all game engines. This system allows the creation of Force Feedback Effect assets that contain properties defining specific haptic effects (Figure 1c). An arbitrary number of channels can be created when creating an asset, and each channel can reproduce a different effect whose pattern is defined by a curve and a list of vibration motors on which that effect will be reproduced. Created Force Feedback Effect can be played on a specific device once or looping, stopped and added to any actor during live gameplay. Effects can be reproduced from a specific source in the video game world, where their intensity will depend on the distance of the source from the player. However, it is not possible to freely define the influence curve of the distance of the player from the source on the intensity of the effect. Unreal Engine offers audio-based vibration as an experimental feature, currently supported only for the PS5 DualSense controller. There are additional features such as trigger feedback, resistance, and vibration, but these are platform-specific and primarily available for consoles and certified developers [11]. Although effects can be previewed in the editor, there is no visual representation of the effect unless a specific effect is opened in the Force Feedback Effect editor. Additionally, there is no playhead to track preview progress.

Interhaptics offers tools like the Haptic Composer and uses HAPS (Haptic Authoring and Playback Standard) files to facilitate the integration of haptic feedback into projects developed with Unity and Unreal Engine. Haptic effects are stored in .haps files, which are general, platform-agnostic, and written in a JSON-based format. These files are created and edited using the external Haptic Composer tool, and they are integrated into projects through APIs and tools provided by Interhaptics SDKs and plugins

for Unity and Unreal Engine. The Interhaptics SDK manages the translation of haptic experiences to the chosen peripheral device [12]. Haptic Composer (Figure 1d) is probably the most comprehensive software tool for creating haptic effects. It supports a wide array of devices and platforms, offers a feature-rich editor and preview for in-app testing, includes haptic presets and enables the translation of audio files to haptic effects. However, it is important to note that this tool is installed and used independently of the game engine.

VibEd is a web-based editor tool made for prototyping of vibrotactile haptic effects for Xbox gamepads, iOS and Android devices. It was developed as a part of research done by Nordvall et al. [4]. This editor visualized the amplitude and durations of signals as a waveform, so users were able to sketch, test and adjust haptic effects (Figure 1e). Effects created with this tool could be tested immediately using a player application on the device and exported into code that can be used in development. The code generated by this editor needs translation for different platforms, requiring dedicated libraries and plugins to reproduce effects on specific devices. No game engines were used to develop the VibEd editor, its plugins, and libraries. Gamepad vibration module is implemented using the Microsoft XNA framework. The VibEd editor, while useful for testing vibration signal ideas more efficiently than writing code, has several limitations identified by designers who participated in the research. It lacks support for explorative sketching and dynamic signals connected to variables. Designers suggested improvements such as incorporating predefined template signals, generating haptic signals from audio and expanding timeline functionality. Specific recommendations included adding copy/cut/paste functions for signals, a grid for time intervals, and a playhead for better timeline navigation. Findings from this paper provide a crucial insight into features that enhance the usability, work efficiency, and user experience of such a tool.

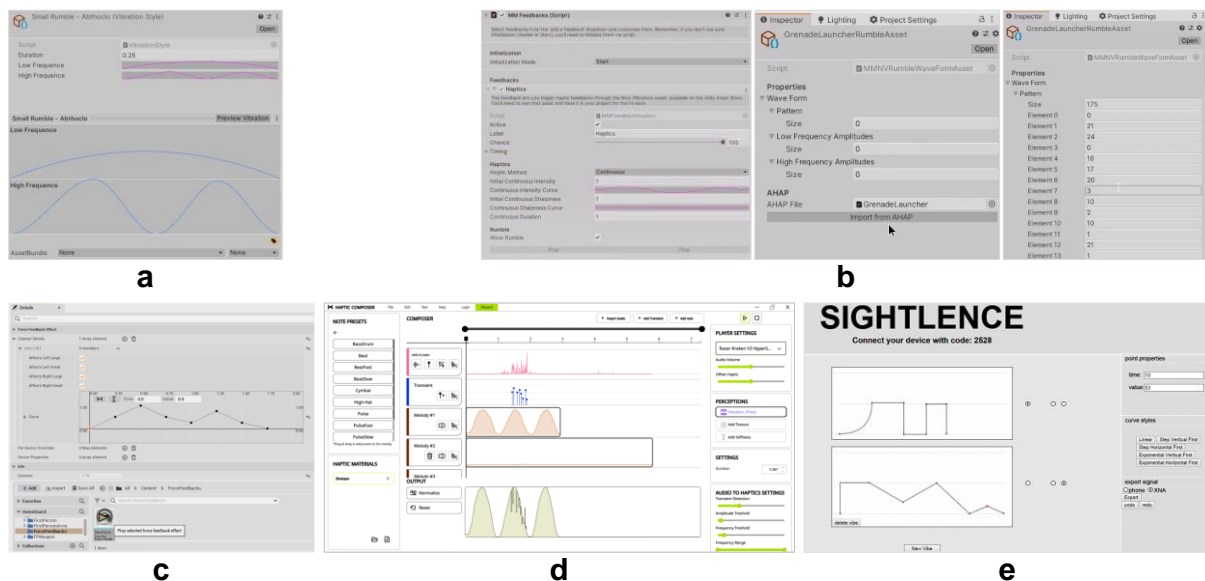


Figure 1. Haptic Effect editor and preview in Haptic package for Unity (a), Continuous haptic feedback and gamepad RumbleWaveFormAsset creation in Nice Vibrations package for Unity (b), Force Feedback Effect editor and preview in Unreal Engine (c), Interhaptics Haptic Composer software tool (d), VibEd web-based editor tool (e)

The program model described in this paper was created with the aim of offering a comprehensive solution based on the experiences of existing solutions, fully integrated within the game engine, and featuring a visual, practical and highly functional preview. During development, the goal was to adopt and enhance the strengths of existing solutions while overcoming their limitations, as much as it is possible in current scenario.

### 3. Design and Implementation

The Unity engine was chosen for the implementation of this solution, because it is scripting-oriented and relies heavily on writing C# code. It allows for extensive editor scripting, enabling developers to create custom editors, tools, and workflows within the Unity Editor. Due to the similarities between the Unity and Godot engines, and Godot's support for the C# programming language, the proposed solution can be easily adapted to work with Godot. Therefore, it can be said that this solution is almost instantly compatible with two of the three most popular game engines currently available. Certainly, its underlying philosophy can also be applied to other game engines, including Unreal Engine, or outside of any game engine, but these would naturally require more significant changes than those needed for Godot adaptation. Finally, the limited range of built-in functionalities related to gamepad vibration within Unity and Godot engines, compared to Unreal Engine, was another motive for this choice.

This solution is designed to be completely generic and to integrate easily with any Unity game. It allows game designers to create haptic effects without programming skills, enabling a clear division of responsibilities within the development team by letting specialized teams focus on designing haptic effects while developers handle their integration and execution in the game. The only prerequisite for using this program model is to implement Unity's Input System package [13] in the project, which has long been the standard for processing user input. Since Unity's Input System is designed to create a common interface for all input devices, this implementation should work for console controllers and any generic controller with vibration motors.

A greater degree of user control requires greater complexity of the program model. During the development of this solution, a usable compromise was sought between the degree of user control and the complexity of the model. The foundation of this program model consists of three main classes, whose functionality, organization and extensibility are improved with several additional classes and enumerations. Class diagram of this program model is shown in Figure 2.

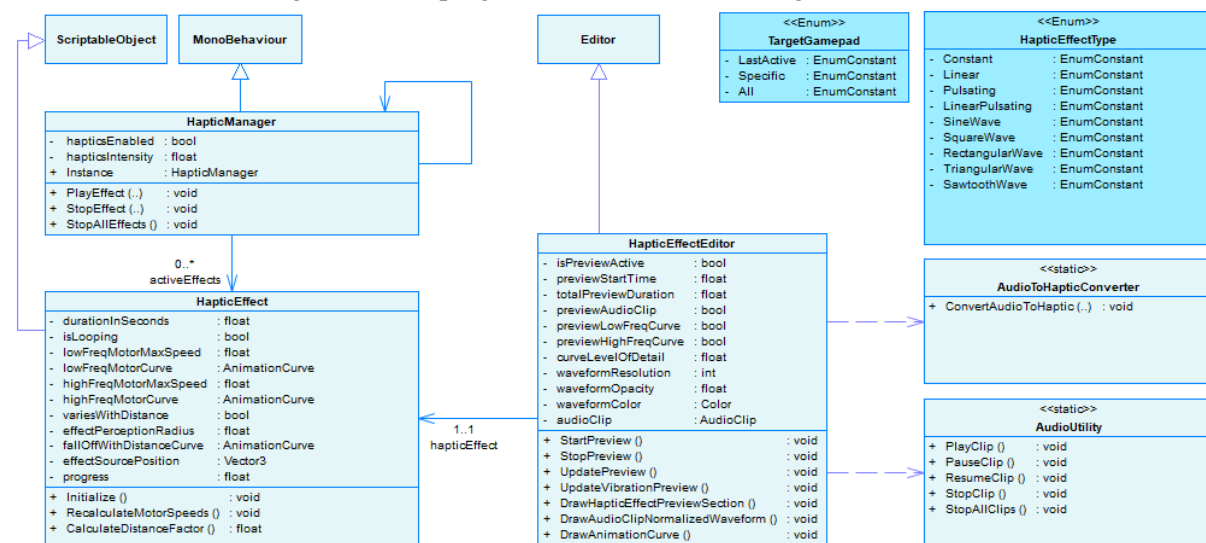


Figure 2. Class diagram of the proposed solution

The **HapticManager** class handles the global configuration and control of haptic effects, balancing and blending simultaneous effects, and integrating with other game systems like audio and camera system. Implemented as a Singleton, it ensures a unique, globally accessible instance that persists across scenes, facilitating high-level haptic feedback management and easy integration into projects. The **HapticManager** can play, pause, resume, and stop one specific or all haptic effects, blending them and sending calculated intensities, i.e. motor speeds to the gamepad. Users can specify the source location and targeted gamepads for each effect. Additionally, it provides methods to programmatically generate predefined haptic effects, such as constant, pulsating, and various wave vibrations, without needing to create **HapticEffect** objects in the Editor. It allows users to adjust or disable haptic feedback through game settings, which is particularly important, because a certain



percentage of players perceive it as distracting, irritating or unpleasant, and excessive exposure to vibrating controllers can, in some cases, result in Hand-arm Vibration Syndrome (HAVS) [14].

The `HapticEffect` class, inheriting from Unity's `ScriptableObject` class, models individual haptic effects. These assets are created in the Unity Editor via the Assets/Create submenu within the Project window, they are lightweight and can be copied into other Unity projects, making them reusable like 3D models and audio effects. Properties of this class, adjustable in the Inspector window, include looping and duration of the effect, motor vibration intensities, and their changes over time and distance using animation curves. By varying and combining these properties, developers can design realistic haptic feedback for diverse game scenarios and events.

The `HapticEffectEditor` class, inheriting from Unity's `Editor` class, defines a custom visual editor for creating and previewing haptic effects (Figure 3). Users can create effects by adjusting animation curves, dragging audio files, or combining both methods. Current algorithm generates curves for gamepad motors by analyzing amplitude changes in the audio clip. The preview section displays animation curves representing the changes of vibration motors' speeds over time and allows playback of haptic effects directly in the editor while the gamepad is connected, without the need to start the project (Figure 4a). A playhead marker tracks effect curve's playback progress, supporting real-time adjustments. When effects are created based on audio clips, a normalized audio waveform is displayed behind the curves, facilitating precise editing and synchronization. Users can toggle the playback of the audio clip or haptic curves during the preview.

During in-game playback, the `HapticManager` calls methods of active `HapticEffect` instances each frame to calculate their individual motor intensities. These calculations are based on the effects' properties, current reproduction progress, and the player's spatial position relative to the effects' sources. The `HapticManager` then blends the calculated intensities of all active effects, determining the final speeds for the gamepad's vibration motors, which are then sent to the gamepad (Figure 4b).

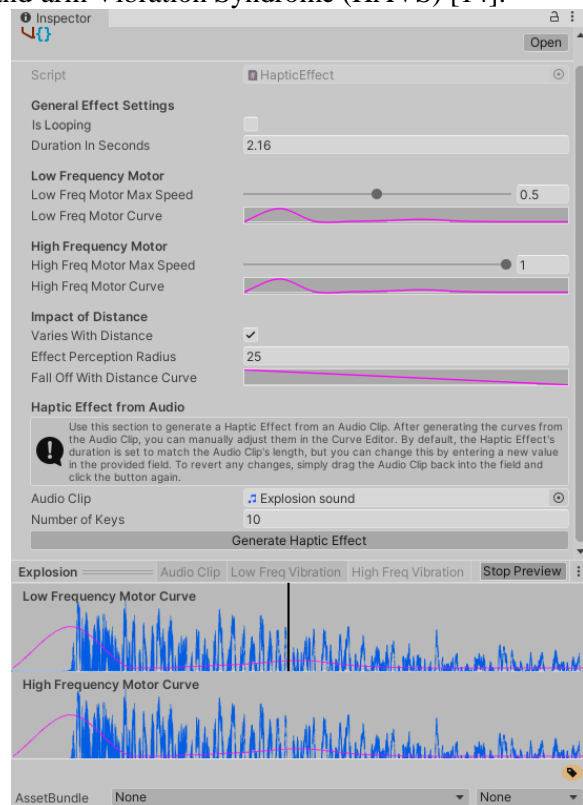


Figure 3. Custom visual editor for design and preview of Haptic Effect object

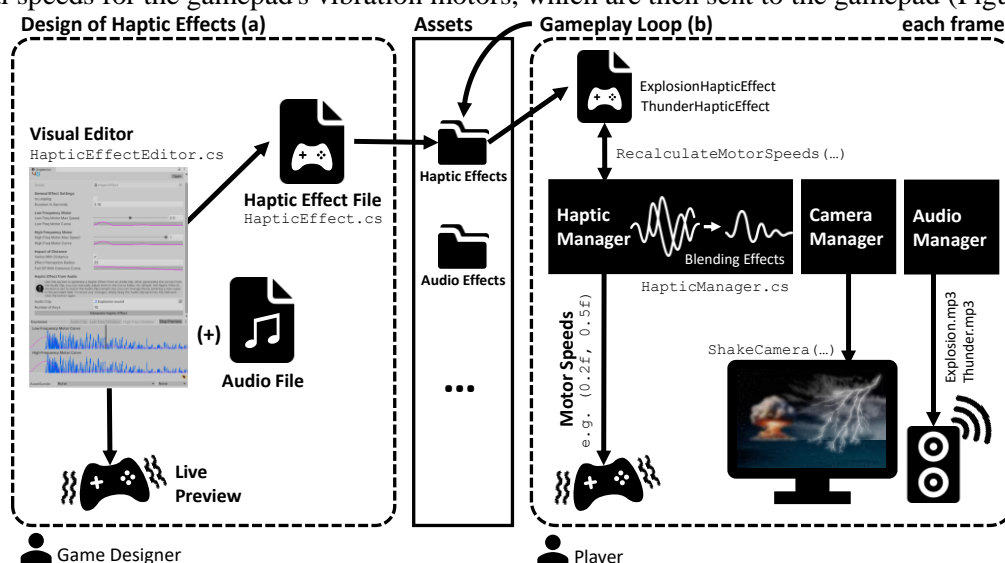


Figure 4. Workflow schema of the proposed solution: Design of Haptic Effects (a), Gameplay Loop (b)

## 4. Limitations and Future Work

This programming model was developed for the generally most common generic gamepad controller with two vibration motors in the handles. It can be adapted to support motors in triggers of the gamepad, such as those found in the Xbox One and Sony's DualSense controllers, but this adaptation requires access to specialized programming support, which is available only to licensed console developers.

Obviously, hardware will put limits on the level of control haptic effects will have over the gamepad motors. The responsiveness of the motors and the rate at which the speed can be updated depend on the hardware and drivers [5]. In general, the quality and level of detail in vibrations are limited due to the use of Eccentric Rotating Mass (ERM) actuator technology, which is prevalent in most gamepad models on the market. Haptic effects created using the solution proposed in this paper will work on next-generation controllers and they will be reproduced much more precisely, sophisticatedly, and tangibly. For instance, motors located in the handles of Sony's latest DualSense controller [15] use Voice Coil Actuator (VCA) technology, which allows for very fast changes in intensity and frequency of vibration. Although these limitations originate from the hardware and not the proposed solution, it is important to consider them when designing haptic effects to minimize their impact on effect reproduction. Additionally, since Input System API calls must be made from the main thread, the timer resolution for playing haptic effects is limited to 1/60 FPS. One possibility to improve playback quality of haptic effects on gamepads is by automatically adjusting the number of breakpoints that can be played by the Input System [7].

The most important area for future work is the improvement of algorithms for generating haptic effects based on audio files. This can be accomplished through a comprehensive research and analysis of the methodologies used by Interhaptics' Haptic Composer and implementation created by Sony for its DualSense controllers [15]. The sound could be represented in the form of a vibration where the left and right audio channels represent the left and right sides of the controller. This approach could be further strengthened by using a machine-learning classifier that is built on expert-labeled datasets [16]. It is also possible to implement the generation of haptic effects based on AHAP files, following the example of the Nice Vibrations package. In both cases, it is necessary to focus on efficient and practical mapping of the generated effects to the animation curves within the editor, in order to maintain the possibility of the user to subsequently change and adjust the effects.

Adapting the philosophy of this program model for other game engines, especially Godot, would be beneficial for expanding its reach and usability in different development environments. Although this version is specifically intended for gamepads, with certain modifications it can be adapted to work for mobile phone vibration systems as well.

## 5. Conclusions

With the increase in the prevalence of gamepads that have the ability to vibrate and the constant improvement of hardware technologies in them, it is up to video game creators to support and more creatively use these capabilities, as haptic feedback plays a crucial role in video game immersion.

Program model described in this paper is designed to be generic and easy to integrate into existing games developed using the Unity engine. It provides users with different approaches to creating reusable haptic effects, a user-friendly graphical interface for editing effects, and a comprehensive, robust preview system directly within the game engine editor. These advantages increase work efficiency and productivity, allowing programmers and game designers to focus on creating realistic, creative, and sophisticated haptic effects that will be reproduced in the game, rather than on implementing logic for processing haptic effects and reproduction of the vibration itself.

Presented solution offers a high degree of flexibility and extensibility, so in the future, it can be adjusted, adapted for other game engines and devices, and synchronized with other feedback systems in the game in order to provide the player with an even more immersive experience. In addition to that, the priority for future work is to improve the algorithms for generating haptic effects, specifically animation curves for the device's vibration motors, based on audio and AHAP files. Furthermore, the limitations of the device's vibration motors should be precisely considered in order to minimize their impact on the reproduction of the created effects.

### Acknowledgment:

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad” (No. 01-3394/1).

### References:

- [1] S. Swink, *Game feel: a game designer's guide to virtual sensation*, 1st ed. Amsterdam ; Boston: Morgan Kaufmann Publishers/Elsevier, 2009.
- [2] U. SöDerström, W. Larsson, M. Lundqvist, O. Norberg, M. Andersson, and T. Mejtoft, “Haptic feedback in first person shooter video games,” in *33rd European Conference on Cognitive Ergonomics*, Kaiserslautern Germany: ACM, Oct. 2022, pp. 1–6. doi: 10.1145/3552327.3552333.
- [3] E. C. Willumsen and M. Jaćević, “A Typology of Rumble,” presented at the DiGRA '19 - Proceedings of the 2019 DiGRA International Conference: Game, Play and the Emerging Ludo-Mix, Sep. 2019.
- [4] M. Nordvall, M. Arvola, E. Boström, H. Danielsson, and T. Overkamp, “VibEd: A Prototyping Tool for Haptic Game Interfaces,” in *iConference 2016 Proceedings*, Philadelphia, USA: iSchools, Mar. 2016. doi: 10.9776/16138.
- [5] Unity Technologies, “Interface IDualMotorRumble | Input System | 1.8.2,” Unity Documentation. Accessed: Jun. 27, 2024. [Online]. Available: <https://docs.unity3d.com/Packages/com.unity.inputsystem@1.8/api/UnityEngine.InputSystem.Haptics.IDualMotorRumble.html>
- [6] Abthoclo, “Haptic - Enhanced Haptic Feedback for Unity,” Haptic Documentation. Accessed: Jun. 27, 2024. [Online]. Available: <https://abthoclo.com/Haptic/Documentation/>
- [7] More Mountains and Lofelt, “Nice Vibrations | Feel Documentation,” Nice Vibrations Documentation. Accessed: Jun. 27, 2024. [Online]. Available: <https://feel-docs.moremountains.com/nice-vibrations.html#gamepads>
- [8] C. Davis, “Haptix - Create & Share Haptic Experiences,” Haptix Official Website. Accessed: Jun. 27, 2024. [Online]. Available: <https://www.haptix.com/>
- [9] Meta, “Haptics Overview | Oculus Developers,” Meta Haptics Studio and Haptics SDK. Accessed: Jun. 27, 2024. [Online]. Available: <https://developer.oculus.com/resources/haptics-overview/>
- [10] Godot community, J. Linietsky, and A. Manzur, “Controllers, gamepads, and joysticks,” Godot Engine Documentation.
- [11] Unreal Engine, “Force Feedback in Unreal Engine | Unreal Engine 5.4 Documentation | Epic Developer Community,” Unreal Engine 5.4 Documentation. Accessed: Jun. 27, 2024. [Online]. Available: [https://dev.epicgames.com/documentation/en-us/unreal-engine/force-feedback-in-unreal-engine?application\\_version=5.3](https://dev.epicgames.com/documentation/en-us/unreal-engine/force-feedback-in-unreal-engine?application_version=5.3)
- [12] Interhaptics, “Haptic Composer - Interhaptics,” Interhaptics Documentation. Accessed: Jul. 03, 2024. [Online]. Available: <https://www.interhaptics.com/doc/haptic-composer/>
- [13] Unity Technologies, “Input System | Input System | 1.8.2,” Unity Documentation. Accessed: Jun. 27, 2024. [Online]. Available: <https://docs.unity3d.com/Packages/com.unity.inputsystem@1.8/manual/index.html>
- [14] A. G. Cleary, H. McKendrick, and J. A. Sills, “Hand-arm vibration syndrome may be associated with prolonged use of vibrating computer games,” *BMJ*, vol. 324, no. 7332, p. 301, Feb. 2002, doi: 10.1136/bmj.324.7332.301a.
- [15] Y. Konishi, “Sony Group Portal - R&D - Stories - Haptics,” Sony Technology. Accessed: Sep. 22, 2023. [Online]. Available: <https://www.sony.com/en/SonyInfo/technology/stories/Haptics/>
- [16] G. Yun, M. Mun, J. Lee, D.-G. Kim, H. Z. Tan, and S. Choi, “Generating Real-Time, Selective, and Multimodal Haptic Effects from Sound for Gaming Experience Enhancement,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–17. doi: 10.1145/3544548.3580787.

# Upgrading Traditional E-Commerce Systems with A Knowledge-Based Recommendation System

Teodora Siljanoska<sup>1</sup>, Natasa Blazeska Tabakovska<sup>1</sup>

<sup>1</sup> Faculty of Information and Communication Technologies -UKLO Bitola, St. Partizanska, nn  
7000 Bitola, North Macedonia

*siljanoska.teodora@uklo.edu.mk; natasa.tabakovska@uklo.edu.mk*

## Abstract:

In today's e-commerce, there is a need for intelligent recommendation systems to help consumers choose products. This paper presents an upgrade to traditional e-commerce systems by implementing a knowledge-based recommendation system. The focus is on collaborative filtering, which uses data on consumer preferences to provide personalized product suggestions. A user-user collaborative filtering algorithm is applied, which groups users according to similarities in their choices and suggests products that are popular among consumers with similar preferences.

## Keywords:

Recommendation system, e-commerce, collaborative filtering, knowledge vectors, cosine similarity.

## 1. Introduction

In today's society, which is deeply connected to the Internet and technologies, information systems play a key role in daily activities, especially in electronic commerce. These systems, known as e-commerce information systems, enable easy and efficient trading of products and services over computer networks, such as the Internet [1]. Due to the huge amount of data they operate with, they are ideal for upgrading with intelligent components [2].

Analyzing the data that is generated on a daily basis is a powerful tool for transforming this data into knowledge, gaining invaluable insights, and discovering patterns and trends previously unknown. The acquired knowledge revolutionizes business operations and is an instrument for companies to survive and remain competitive. The benefits that businesses receive from the transformation of data into knowledge are multiple: improved decision-making guided by data and knowledge, better user experience, improved sales and optimization of processes in the organization, that is, increased efficiency and productivity.

For those reasons, the main emphasis in this paper is on the possibilities of upgrading e-commerce systems with a knowledge-based recommendation system and the way to implement such type of upgrade. The focus of the paper is to synthesize new knowledge and identify techniques for upgrading classic e-commerce systems with intelligent components through the analysis of relevant literature and examples, with a special emphasis on knowledge-based recommendation systems. After the introductory part, an analysis of relevant literature from the researched area is made, from where the main idea comes, the implementation of a proposal - a solution that includes upgrading the Reusable system for e-commerce with a recommendation system based on knowledge, which is presented in the third part of the paper. In the final part, concluding observations and future recommendations for further advancement and improvement of the proposed implementation are given.

## 2. Previous work

A group of researchers highlight two factors as the main ones that determine the future survival and success of organizations, namely electronic commerce [3] and knowledge obtained from consumer data

[4]. They encourage the use of e-commerce and the Internet as a platform for accessing and gathering consumer knowledge.

E-commerce systems abound with diverse data on products, consumers, their orders that can serve to generate knowledge and extend it with intelligent components [5]. Analysis of this data and its transformation into business knowledge will allow a better understanding of customers by analyzing their purchasing preferences. This information can be used to improve marketing campaigns, user experience or product range, but very importantly it can be used for personalized user recommendations. Knowing customers allows them to receive individual product recommendations based on their previous purchases or the purchases of other customers with similar preferences. It leads to customer satisfaction and increased sales by providing relevant products at the right time. Equally, decision makers can determine which of their products perform best in inventory management and can find out whether their prices are attractive to customers.

However, subjective measures of user experience from information systems do not always emphasize the importance of cognitive and "emotional" perceptions, especially in the context of online shopping [6] [7] [8]

In that context, ontologies play a key role, which are increasingly present nowadays. Their number, complexity and the domains they model are constantly increasing significantly. Conversion of databases into ontology is one of the key areas of research in the field of Data Science and Knowledge Science. Ontologies, as structures defining relationships between concepts, are acknowledged in this paper as integral to knowledge representation in recommendation systems. In this work, knowledge vectors are applied as a type of ontology to represent consumer preferences and product relationships.

When data acquires semantic meaning, knowledge can be generated from it in two ways: by creating knowledge graphs or by creating knowledge vectors [9]. Knowledge graphs can often become inefficient and challenging to manage as data volume grows, impacting scalability and performance. For this reason, this paper focuses on knowledge vectors, which represent knowledge mathematically as multidimensional vectors, with each dimension indicating a specific attribute or meaning. By using knowledge vectors, consumer preferences and product relationships are quantified efficiently in the background, making them well-suited to the needs of this recommendation system.

According to Tarus, Niu and Mustafa, in recent years, knowledge vectors are increasingly applied to recommendation systems in order to overcome the problem of lack of consumer-product interaction and to represent the large number of attributes in one ontological structure [10].

Recommendation systems are grouped into three basic classes [11]: collaborative filtering: analyzes the activities of the user and other users with similar interests to predict what he would like and suggest appropriate products; content-based filtering: relies on keywords from product descriptions and content related to the user profile to make recommendations; hybrid filtering: combines techniques from the previous two methods for a wider range of recommendations. Furthermore, collaborative filtering is divided into three subgroups: user-user: based on preferences of similar users; product-product: recommends products similar to previously selected products by the user; combined systems: use data about user preferences for specific products.

### **3. Upgrade of an e-commerce system with a knowledge-based recommendation system**

Reusable e-commerce information system is a system consisting of integrated systems: Airtable, Ongoing, WooCommerce, Refurbed and the Swedish accounting system - Fortnox system. Serving as the central hub, it handles data and information about the products and the orders made and presents them on an intuitive Graphical User Interface (GUI) through which the products are managed before they are published for sale on different platforms and with the orders made. In other words, the system provides an opportunity for quick and easy publication of products for sale on the Reusable or Refurbed site, as well as review, monitoring and editing of orders made by consumers.

Building on methodologies for transforming consumer data into valuable insights, this enhanced e-commerce system goes beyond conventional data processing, converting operational data into meaningful knowledge that greatly increases its practical value. By integrating an intelligent

component, the system now supports knowledge-based recommendations, offering tailored and more effective suggestions for consumers, ultimately boosting both user experience and system performance.

Ontology plays a pivotal role in this context as a foundational framework that facilitates extracting meaningful knowledge from data. It provides a structured representation of concepts and relationships within the data domain, enabling the organization and interpretation of complex information. The ontological framework is a key source of ideas for generating knowledge vectors, effectively encapsulating consumer preferences and product relationships in a mathematically representable format. Knowledge vectors are utilized to represent consumer preferences and facilitate the creation of personalized recommendations based on the semantic meaning derived from data.

To implement a recommendation system that is based on the extracted knowledge about the preferred products by consumers, knowledge vectors were used, which according to Guntars are easy to group by similarity [12], thus giving each consumer a suitable proposal.

Based on the primary classification, the specific recommendation system belongs to the group of collaborative filtering systems and that of the user-user class. Namely, for a certain consumer/user, the system provides a proposal - products based on the products preferred by users with similar preferences.

This recommendation system follows several steps:

- **Data loading, preprocessing, and transformation:** Relevant data, including product listings, consumer orders, and itemized purchases, is loaded and preprocessed using the SQLAlchemy library. Data cleaning is performed with NumPy to ensure consistency and relevance, after which order data is merged with item details, creating a refined dataset for analysis.
- **Construction of consumer-product interaction matrix:** an interaction matrix is built, where the rows represent the consumers and the columns represent the products. In the intersection of the row and the column is the ordered quantity of the product by the consumer.
- **Extraction of knowledge from data and knowledge vector creation:** knowledge vectors are generated in the background from the consumer-product interaction matrix. For each user a vector is created that represents their preferences and for each product a vector is created that represents its "occurrence" in orders or how many it is often commissioned and by whom.

- **User knowledge vector:** This vector reflects a user's interactions across products. For example, the vector for user ana@example.com is:

$$U_{ana} = [3, 0, 5, 2, 1], \quad (1)$$

representing her engagement levels:

1. **Product 3:** High engagement:

$$U_{ana,3} = 5 \quad (2)$$

2. **Product 5:** Minimal interaction:

$$U_{ana,5} = 1 \quad (3)$$

3. **Product 2:** No interaction:

$$U_{ana,2} = 0 \quad (4)$$

- **Product knowledge vector:** constructed and represented analogously to user vectors, these vectors quantitatively capture engagement levels across users for each specific product.
- **Application of the collaborative filtering algorithm:** on the obtained knowledge vectors, the collaborative filtering algorithm is applied through:
  - **Calculation of cosine similarity between consumers:** using the Scikit-learn library, the similarity of the obtained knowledge vectors is calculated according to the cosine similarity (with a range of values from -1 to 1) that calculates the angle between two vectors [13]. A lower value simultaneously means a lower similarity in preferences and vice versa.
  - **Identification of similar consumers:** for the given consumer, the most similar ones according to the cosine similarity are found.
  - **Calculation of average product quantity:** purchased by consumers with similar preferences, the average quantity purchased is calculated.

- Exclusion of previously purchased products: products already purchased by the target consumer are excluded, as they are less likely to be repurchased.
- Generation of final recommendations: the system provides recommendations based on the top n products purchased by consumers with similar preferences that the target consumer has not yet bought.

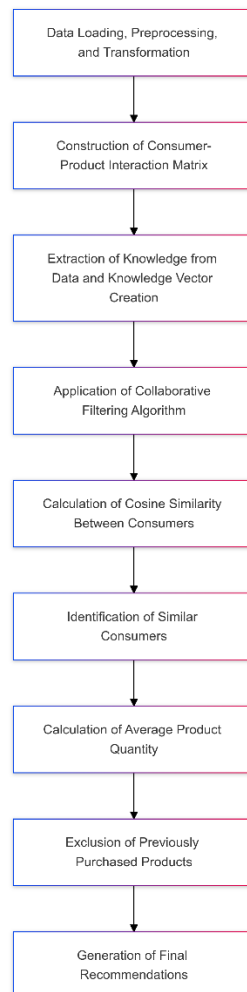


Figure 1: Workflow of the proposed recommendation system

Source: From the Authors

The proposed system is fully functional and allows the generation of product recommendations based on consumers with similar preferences. So for example, for a consumer with email address ana@example.com, product recommendations will be generated based on products consumed by consumers most similar to this consumer. By activating the appropriate system and entering the email address as an input parameter, the system generates recommendations (Figure 2).

```

PS C:\Users\teodo> python knowledge_based_recommendation.py ana@example.com
[{"instance_name": "Microsoft Lumia 950 | 32 GB | black", "sku": "10055251-223"}, {"instance_name": "REFISHED - SPORTY BAG #FISH yellow | yellow dollar | size S", "sku": "refished_41261125501121"}, {"instance_name": "Microsoft Surface Duo 2 | 8 GB | 128 GB | Single-SIM | black", "sku": "Microsoft Surface Duo 2 | 8 GB | 128 GB | Single-SIM | black B"}, {"instance_name": "LG G8s ThinQ | 128 GB | Mirror Black", "sku": "LG G8s ThinQ | 128 GB | Mirror Black C"}, {"instance_name": "Fujitsu Lifebook E546 | i5-6300U | 14\" | 16 GB | 512 GB SSD | FHD | DVD-RW | Webcam | Win 10 Pro | 4G | DE", "sku": "00453-QZ-G-C-12"}, {"instance_name": "CAT S41 | 32 GB | Single-SIM | black", "sku": "CAT S41 | 32 GB | Single-SIM | black C"}, {"instance_name": "CAT B35 | 4 GB | Single-SIM | black", "sku": "CAT B35 | 4 GB | Single-SIM | black B"}, {"instance_name": "Sony Xperia 10 III | 6 GB | 128 GB | Dual-SIM | black", "sku": "SO-X08T52-128-DS-blk-nw"}, {"instance_name": "Samsung Galaxy S20+ | 12 GB | 128 GB | 5G | Dual-SIM | cosmic black", "sku": "35642"}, {"instance_name": "iPhone 13 | 256 GB | Single-SIM | blue", "sku": "IPHONE 13 256GB - AZUL - C"}]

```

Figure 2: Application of the recommendation system

Source: From the Authors

Executing the Python script is from the location where it is placed via the "python" command. The first argument is the name of the script to be executed (`recommendations_test.py`), while the second argument is the email address of the specific consumer for whom recommendations should be generated (`ana@example.com`) (Figure 2). By applying collaborative filtering and taking into account the preferences and products ordered by users most similar to this consumer, the system results in 10 products along with their specifications. The specifications include the product name (instance name) and the product code (sku). Recommended products for this consumer include several smartphones, such as CAT S41, CAT B35, Microsoft Lumia, Sony Xperia 10 III, Samsung Galaxy S20+, LG G8s ThinQ and iPhone 13, a laptop backpack (Upcycling Deluxe), as well as two laptops ( HP Pro x2 612 G2 and Fujitsu Lifebook E546).

#### 4. Contribution

This paper contributes to the advancement of e-commerce systems by introducing a sophisticated knowledge-based recommendation framework, centered on user-user collaborative filtering. The system leverages knowledge vectors and cosine similarity to identify patterns in consumer behavior, facilitating the generation of precise, personalized product recommendations. By transforming raw transactional data into actionable insights, the proposed model enhances user satisfaction, drives targeted marketing efforts, and improves overall system efficiency. The integration of intelligent components within traditional e-commerce architectures paves the way for more informed decision-making and optimized consumer experiences.

#### 5. Conclusion

The knowledge-based recommendation system successfully upgrades the classic Reusable e-commerce system with an intelligent component, enabling personalized product suggestions. By applying user-to-user collaborative filtering, the user experience is improved through recommendations based on similarities between consumers. More precisely, through the application of knowledge vectors and the calculation of cosine similarity, similar consumers are identified and product recommendations are issued. Focusing on the preferences of similar users while excluding already purchased products enables relevant and targeted recommendations that will meet user preferences. This approach significantly contributes to improving the efficiency of e-commerce, optimizing the product selection process for each individual user.

Future upgrades to the system could also include product-to-product recommendations, where similarity is determined by the characteristics of the products themselves. This approach would offer an even deeper level of personalization, enabling recommendations for products that the user has not purchased, but are similar to those they have been considering. Such a strategy will not only increase user satisfaction, but also significantly contribute to long-term sales growth and customer loyalty, setting the e-commerce system at a higher level of efficiency and competitiveness.

#### References:

- [1] R. Fulbright and A. Dudezerte, "Knowledge Management 2.0: Organizational," Hershey: Information Science Reference., 2019.
- [2] B. Gaspar, "Collection, assessment, storage, processing, and use of data and information," University Center, Varazdin, 2020.
- [3] A. Gupta, B. Su and Z. Walter, "Risk profile and consumer shopping behavior in electronic and traditional channels," *Decision Support Systems*, 38(3), 2004.
- [4] M. Ming-Tai Tsai and C. Shih, "The Impact of Marketing Knowledge among Managers on Marketing Capabilities and Business Performance," *The International Journal of Management*, pp. 117-119, 2004.
- [5] I. Boughzala and A. Dudezert, *Knowledge Management 2.0: Organizational Models and Enterprise Strategies*, Hershey: Information Science Reference, 2019, pp. 34-38.
- [6] D. Gefen, E. Karahanna and D. W. Straub, "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly*, vol. 27, no. 1, pp. 51-90, 2003.



- [7] M. Koufaris, "Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior," *Information Systems Research*, vol. 13, no. 2, pp. 205-223, 2002.
- [8] P. A. Pavlou and D. Gefen, "Building Effective Online Marketplaces with Institution-Based Trust," *Information Systems Research*, vol. 15, no. 1, pp. 37-59, 2004.
- [9] V. Stanko, *Information Technologies for Knowledge Management*, Varazdin: Faculty of Organization and Informatics, University of Zagreb, 2022, pp. 46-48.
- [10] J. Tarus, Z. Niu and G. Mustafa, "Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning," *Artificial Intelligence Review*, vol. 50, pp. 21-48, 2018.
- [11] N. Blazheska-Tabakovska, *Knowledge based systems*, Bitola: Macedonian science society - Bitola, 2023, pp. 138-139.
- [12] B. Guntars, "Mapping between Relational Databases and OWL Ontologies: an Example," *Scientific Papers*, vol. 756, pp. 203-205, 2020.
- [13] A. Verheij, *Supporting knowledge management within Supporting knowledge management within municipalities and CJG's with Web 2.0 technologies*, Den Haag: Faculty of Technology, Policy and Management, Executed at Alares, Den Haag, 2021, pp. 120-124.

# Internet of Things in the Automotive Industry: A Review

Maša Knežević<sup>1,2</sup>, Goran Sladić<sup>2</sup>, Branko Markoski<sup>2</sup> and Aleksandra Stojkov Lončarski<sup>3</sup>

<sup>1</sup> Office for Information Technologies and e-Government, Belgrade, Serbia

<sup>2</sup> University of Novi Sad, Faculty of Technical Science, Novi Sad, Serbia

<sup>3</sup> University Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

masa.knezevic@ite.gov.rs; sladicg@uns.ac.rs; markoni@uns.ac.rs; aleksandra.stojkov@tfzr.rs;

## Abstract:

In this paper, the authors highlight the fundamental elements of the automotive industry and the statistics on the demand for cars. Through a review of various literature, the authors emphasize the most important segments within the application of the Internet of Things in the automotive industry. This paper presents the basic concept and application of the Internet of Things in various areas, facilitating rapid and efficient implementation. The aim of this paper is to present literature where specific completed projects are showcased, clearly demonstrating the application of the Internet of Things in the automotive industry. This paper presents solutions that aid in car parking in urban areas, as well as the navigation of cars and other vehicles via GPS.

## Keywords:

Automobiles, Automation, sensors, Internet of Things, Automotive industry

## 1. Introduction

The Internet of Things (*IoT*) is a relatively new technology with numerous innovations, initially conceived by Kevin Ashton in 1999 for a sensor project. It refers to a network of physical objects equipped with sensors, software, and other technologies to exchange data with other devices and systems over the Internet. These objects range from everyday household items to advanced industrial tools. McKinney defined IoT as the integration of sensors and actuators in physical objects connected via wired or wireless networks, often using the Internet Protocol (*IP*). [1]

Technology is an indispensable part of our lives, constantly transforming us and enabling access to an increasing number of achievements. In that light, the relatively new phenomenon, the Internet of Things, or machine-to-machine (*M2M*) technology, opens the door to numerous possibilities and promises. IoT is defined as a network where smart machines collectively collect data, exchange information, process it, and take actions, autonomously and in collaboration with others. This technology brings about a revolution in how we interact with our environment, allowing machines to autonomously respond to data and the surrounding environment. One of the key elements driving the development of IoT is the use of artificial intelligence (*AI*), known as Intelligent IoT. This technology enables machines to mimic human thinking and decision-making, thereby improving the quality of services they provide. Through the use of AI, IoT becomes capable of providing more personalized and accurate services, contributing to higher levels of reliability and user satisfaction. Because of these innovations, IoT becomes a key player in our lives, transforming the way we perform everyday activities and tasks. From smart homes to smart cities, IoT opens up unlimited possibilities for improving efficiency and comfort in life. IoT represents a revolution in how technology is integrated into our lives. With an increasing number of smart devices interconnected and communicating with each other, the future that IoT brings promises to be increasingly intelligent, efficient, and personalized. [8]

The introduction of the Internet of Things technology into the automotive industry represents a revolutionary step towards creating smarter, more connected, and safer vehicles [2]. In the early 2010s, the automotive community recognized the potential of IoT for remotely monitoring and addressing vehicle defects, resulting in reduced issues and improved vehicle quality. The integration of Dedicated Short-Range Communication (*DSRC*) devices into the Vehicle-to-Everything (*V2X*) communication system in 2012 further enhanced IoT performance in vehicles, enabling more efficient data exchange

between vehicles and infrastructure [3]. At the same time, research has shown that IoT can be used to address key transportation challenges, such as monitoring road conditions and detecting irregularities, further contributing to improved traffic safety and efficiency.

The first major step in commercializing IoT technology in automobiles was the introduction of the Toyota Mirai in 2016, as the first mass-produced car to utilize IoT for performance and functionality enhancement. This event marked a milestone in the development of the automotive industry, demonstrating the power of IoT in transforming traditional vehicles into smart, connected entities.

The rapid growth of IoT technology in the automotive industry has not gone unnoticed in the market. Analysts have predicted that by 2020, around 250 million vehicles will be equipped with IoT technology, while the market value of IoT is set to skyrocket from \$20 billion in 2016 to \$100 billion in 2023. This rapid growth underscores the increasing importance of IoT in improving transportation, facilitating travel, and enhancing road safety. [4]

Through real-time vehicle performance monitoring and fast detection of irregularities, IoT technology not only improves the efficiency of smart vehicle production but also enhances the driving experience by enabling better information sharing among drivers, reducing travel disruptions, and decreasing the number of traffic accidents. The automotive industry, along with related sectors, has become a leader in implementing IoT technology, setting standards for the future of connected mobility. [5,6,7]

The authors have divided this paper into several parts. The first segment of the paper is a research problem in automotive industry in general. The second segment is about basic concept of the Internet of Things. And the third segment is literature review of Internet of Things in the Automotive industry. The final part is conclusion which is the part for future research.

## 2. Research problem

The automotive industry produced over 93.5 million vehicles in the previous year of 2023. This fact indicates exponential growth in the automotive industry compared to previous years of vehicle production. The leading country in the world in vehicle production in the previous year of 2023 is China, which produced over 30 million vehicles. The second-largest vehicle-producing country is the USA with 10.6 million vehicles, while Japan ranks third with 8 million vehicles produced. Innovations in production contribute to facilitating people's daily lives as well as technical and technological advancement. According to research, customers today seek futuristic products with which they can connect and synchronize to fully meet their needs. The emergence of new functionalities in the automotive industry provided by IoT results in customer satisfaction and increased sales. [9]

Currently, the automotive industry requires automated and diverse multipurpose applications to address current challenges in the transportation sector. Self-driving cars, trucks, and other vehicles are now present and actively used in certain countries worldwide. The development of the automotive industry does not stop there but aims to provide the best possible experience for its customers. It is in this segment that synergy between the automotive industry and IoT occurs. [10]

This paper will showcase examples of synergy between IoT and the automotive industry, while also highlighting potential issues that can be addressed in the near future.

## 3. Concept of the Internet of Things

In the near future, the number of internet-connected things will vastly exceed the number of people. Objects around us will be linked to the internet in various ways, enabling machines to communicate with each other instead of people. Various technologies such as mobile networks, satellites, WiFi, Bluetooth, and others will become part of the IoT, facilitating machine-to-machine communication. The key aspect of the IoT vision involves integrating smart devices with intelligence, sensors, and remote control capabilities. The Internet of Things is an advanced concept in the world of information technology that has been developing over the past decade. This technology enables a global network of interconnected physical objects, facilitating their communication with each other and with humans. IoT transforms the way objects in the world operate, allowing them to sense their environment, communicate, and react without human intervention. At the core of IoT are sensors and actuators

embedded in various items, generating large amounts of data for analysis. The Internet of Things was initially inspired by technologies like Radio Frequency Identification (*RFID*) and NFC, which enable the discovery of information about tagged objects over the internet. RFID is a key technology for building IoT [11]. IoT integrates physical objects into the cyber world using sensors, radio frequency, and other technologies for data collection and analysis. This technological revolution spans various aspects of the IT industry, from computing and communication networks to new technologies such as remote data collection and analysis.

The application of the Internet of Things has wide usage in solving various industrial problems. Combining IoT with other techniques and technologies leads to significant advancements in different domains. For example, to reduce high electricity consumption in daily households, IoT sensor networks combined with image processing are applied. This approach can achieve the identification of human activities through image processing and adjust energy supply as needed with the help of IoT.

Authors [12] state that a big data technologies such as NoSQL MongoDB, Apache Kafka, and Apache Storm play a crucial role in processing and storing data in the manufacturing industry. Apache Kafka provides a scalable infrastructure for real-time data processing, essential for fast and efficient analysis of large datasets. In healthcare, sensors connected to the Internet of Things are increasingly used to gather patient data. These data are often stored in systems like Apache Kafka and MongoDB, enabling continuous monitoring of health parameters and real-time data analysis. This has the potential to enhance diagnostics, patient monitoring, and personalized medical care. Internet of Things also brings significant changes in data security, transmission, fault tolerance, and transparency. This technology emphasizes decentralization and openness, enabling a wide range of applications across various industrial and business sectors. In agriculture, IoT technology is applied for supply chain management using platforms such as Hyperledger Fabric and Ethereum. These platforms enable precise tracking of products throughout the supply chain, improving efficiency and transparency in distribution and resource management. Integrating IoT technology requires careful planning and implementation, ensuring adequate data protection and compliance with privacy regulations. It is crucial to securely transmit and store data, while strictly adhering to legal and regulatory requirements to ensure effective and sustainable deployment of IoT solutions across all industrial contexts.[13,14]

One of the main issues with the Internet of Things is its vastness and the absence of a unified architecture [15]. For IoT to function effectively, it requires a combination of various technologies such as sensors, networks, communications, and computing. Different researchers, authors, and practitioners have proposed various architectures and models of IoT to address these challenges. The Internet of Things is bringing significant technological changes into our lives, facilitating them through various applications in medicine, industry, transportation, education, and other fields. Despite offering numerous benefits, IoT faces challenges in governance and implementation, such as the lack of a global definition, the need for architectural-level standardization, interoperability among different technologies, and the necessity for protocol standardization to enhance global management.[16]

#### **4. Internet of Things in the Automotive Industry**

Recent technology research predicts that in the coming years, it will be possible to produce safer, smarter, zero-emission, and more energy-efficient vehicles thanks to advanced technologies such as wireless sensors, embedded systems, smart networks, lightweight materials, efficient engines, cloud computing, batteries, fuel cells, and intelligent converters. In addition to automotive technologies, other critical factors include road infrastructure, vehicle maintenance, driver attitudes, and regulations, which must work together to preserve nature and support clean environments and sustainable mobility [17]. The semiconductor industry is increasingly focusing on the automotive and industrial markets, particularly in developing intelligent IoT applications for workshop automation and network structures [18]. The complexity of software in modern vehicles requires rapid implementation of new features and improvements, posing a challenge for manufacturers due to the limited capabilities of current software updates in service centers [19]. Kirk and colleagues explore the possibilities of IoT technology, emphasizing device control via smartphones and the rapid development of connected devices in autonomous driving and transportation. They also propose an IoT-based system for dismantling old vehicles, while Liu and colleagues highlight the rapid advancement of IoT and its impact on the

automotive industry. This reflects excitement about integrating IoT into daily life, with a focus on smartphones and the continuous advancement of connected technologies.[20]

Intelligent transport represents a global trend. European Union member states are collectively working on the development of intelligent transport. The authors conduct research contributing to the knowledge about the global traffic management agenda. Vehicle management during parking processes in urban environments is part of a trend leading to technological advancement. The Internet of Things enables access to data about physical objects such as parking spaces through embedded communication technologies via IP-based networks. This allows migration towards a fully IP environment in a later stage, where full automation of service-oriented use cases reveals the potential benefits of the proposed framework in this paper. The authors explore horizontal contributions to various research areas such as Internet of Things, Sensor Services, Business Modeling, Computer Search, Big Data Analytics, Real-Time Processing, and implementation of smart parking applications on cloud platforms. Additionally, the focus is on innovations in security and privacy, crucial for access management at the sensor, service, and application levels. The paper considers four key dimensions of issues in parking systems. The first dimension is the interoperability of information and services in a globally connected continuum, where middleware has been developed to facilitate communication with parking sensors. The second dimension addresses service optimization through cloud brokers to meet interdependency and security compliance requirements. The third dimension explores the use of machine learning for processing real-time data from parking sensors, while the fourth dimension develops a domain-specific language for modeling business processes and dynamic pricing of parking services in cloud-supported middleware.[21]

Car-2-X technology represents a key element in achieving cleaner, safer, and more efficient traffic by enabling vehicles to wirelessly exchange information among themselves and with infrastructure. However, traditional wireless technologies like IEEE 802.11p, often used in Car-2-X applications, are characterized by narrow bandwidth and susceptibility to multipath fading and inter-user interference. Introducing ultra-wideband (UWB) technology offers a potential solution to these limitations. UWB radios transmit signals over a wider bandwidth ( $\geq 500\text{MHz}$ ), resulting in higher immunity to multipath fading, reduced interference, high data transmission rates, and exceptional time-domain resolution [22]. This technology enables precise distance estimation between vehicles, crucial in scenarios such as cooperative platooning. To ensure reliable communication in all environments, a new scheme has been developed to dynamically adjust UWB physical layer (PHY) settings in real-time. This scheme optimizes communication performance and accurately determines distances between vehicles. The approach begins with an overview of the adaptive algorithm, detailed analysis of PHY setting performance, and real-time environment characterization using channel impulse response information. Evaluation of the proposed algorithm confirms its efficiency and reliability in practical Car-2-X technology applications.[23]

The rapid advancement of technology and the increasing adoption of the Internet of Things have facilitated the creation of sophisticated methods for tracking mobile assets like cars and office equipment in real-time. Key technologies driving this progress include Wireless Sensor Networks (WSN), RFID and the Global Positioning System (GPS) [24]. IoT technology allows for dynamic monitoring of the location and status of various assets, such as cars, enabling real-time oversight and control of valuable resources via the internet. According to [25] and [26], IoT represents an innovative perspective and is critically important for effectively tracking valuable assets that are constantly in motion. However, the deployment of GPS in unstable environments remains a challenge due to its crucial role for users. The concept of the Internet of Things introduces opportunities for an open network environment where objects or things possess autonomous or intelligent entities. These entities collaborate seamlessly with other objects and operate independently based on their respective scenarios [27]. In the context of vehicle tracking, the GPS system facilitates straightforward navigation and monitoring of any car [26]. Within the IoT paradigm, the GPS module serves as a vital component. In practical terms, a tracked vehicle receives data via the network from sensors or RFID tags. The GPS transmits data in bytes that contain precise location details of the object, such as a car, including timestamp, longitude, latitude, and altitude. These details are essential for real-time tracking and management of vehicles.[28]

## 5. Conclusions

The Internet of Things (IoT) is revolutionizing the automotive industry, bringing numerous innovations and improvements in safety, efficiency, and user experience. The application of IoT technologies in smart cars, fleet management, and security systems has shown significant potential in transforming traditional vehicles into intelligent and connected machines. IoT enables the development of autonomous vehicles and connected cars, providing drivers with advanced functionalities such as self-driving, real-time navigation, and personalized user experiences. IoT technologies enhance fleet management efficiency through real-time tracking, predictive maintenance, and route optimization, leading to cost reduction and increased productivity. The integration of IoT in the automotive industry brings advanced safety functionalities, including accident prevention systems and cybersecurity, thus increasing road safety.

IoT brings numerous benefits, such as improved safety, efficiency, and user experience. However, it also faces challenges, including data privacy concerns, integration of various technologies, and the need for developing an appropriate regulatory framework. Trends indicate further integration of IoT technologies in the automotive industry, with a focus on developing fully autonomous vehicles, enhancing safety systems, and personalizing the user experience. Innovations such as V2X (vehicle-to-everything) communication and advanced driver assistance systems (ADAS) are expected to become standard in the future.

IoT is a key factor in the evolution of the automotive industry, bringing numerous benefits to manufacturers, drivers, and society as a whole. Although there are challenges to overcome, the advantages IoT offers in terms of safety, efficiency, and user satisfaction outweigh these obstacles. Further research and development in this field promise even greater innovations and advancements, making the future of the automotive industry exciting and transformative.

### Acknowledgment:

The Author Maša Knežević would like to thank The Office for Information Technology and eGovernment for their support.

### References:

- [1] Fatima, S., Haleem, A., Bahl, S., Javaid, M., Mahla, S.K. and Singh, S., 2021. Exploring the significant applications of Internet of Things (IoT) with 3D printing using advanced materials in medical field. *Materials Today: Proceedings*, 45, pp.4844-4851.
- [2] Kamil, I.A. and Ogundoyin, S.O., 2019. Lightweight privacy-preserving power injection and communication over vehicular networks and 5G smart grid slice with provable security. *Internet of Things*, 8, p.100116.
- [3] Ninan, S., Gangula, B., von Alten, M. and Sniderman, B., 2015. Who owns the road? The IoT-connected car of today—and tomorrow. Deloitte University Press, August, 18, p.2015.
- [4] Guo, H., Zhou, X., Liu, J. and Zhang, Y., 2022. Vehicular intelligence in 6G: Networking, communications, and computing. *Vehicular Communications*, 33, p.100399.
- [5] Eris, O., Drury, J.L. and Ercolini, D., 2021. A collaboration-centric taxonomy of the Internet of Things: implications for awareness support. *Internet of Things*, 15, p.100403.
- [6] Ashouri, M., Davidsson, P. and Spalazzese, R., 2021. Quality attributes in edge computing for the Internet of Things: A systematic mapping study. *Internet of Things*, 13, p.100346.
- [7] Pourrahmani, H., Yavarinasab, A., Zahedi, R., Gharehghani, A., Mohammadi, M.H. and Bastani, P., 2022. The applications of Internet of Things in the automotive industry: A review of the batteries, fuel cells, and engines. *Internet of Things*, 19, p.100579.
- [8] Chen, Y.K., 2012, January. Challenges and opportunities of internet of things. In 17th Asia and South Pacific design automation conference (pp. 383-388). IEEE..
- [9] Zhao, M., Fang, Y. and Dai, D., 2023. Forecast of the evolution trend of total vehicle sales and power structure of China under different scenarios. *Sustainability*, 15(5), p.3985.
- [10] Fraga-Lamas, P. and Fernández-Caramés, T.M., 2019. A review on blockchain technologies for an advanced and cyber-resilient automotive industry. *IEEE access*, 7, pp.17578-17598.

- [11] Mezzanotte, P., Palazzi, V., Alimenti, F. and Roselli, L., 2021. Innovative RFID sensors for Internet of Things applications. *IEEE Journal of Microwaves*, 1(1), pp.55-65.
- [12] Ge, M., Bangui, H. and Buhnova, B., 2018. Big data for internet of things: a survey. *Future generation computer systems*, 87, pp.601-614.
- [13] Omolara, A.E., Alabdulatif, A., Abiodun, O.I., Alawida, M., Alabdulatif, A. and Arshad, H., 2022. The internet of things security: A survey encompassing unexplored areas and new insights. *Computers & Security*, 112, p.102494.
- [14] Khodadadi, F., Dastjerdi, A.V. and Buyya, R., 2016. Internet of things: an overview. *Internet of things*, pp.3-27.
- [15] Ray, P.P., 2018. A survey on Internet of Things architectures. *Journal of King Saud University-Computer and Information Sciences*, 30(3), pp.291-319.
- [16] Malik, P.K., Sharma, R., Singh, R., Gehlot, A., Satapathy, S.C., Alnumay, W.S., Pelusi, D., Ghosh, U. and Nayak, J., 2021. Industrial Internet of Things and its applications in industry 4.0: State of the art. *Computer Communications*, 166, pp.125-139.
- [17] I.B. Aris, R.K.Z. Sahbusdin, A.F.M. Amin, Impacts of IoT and big data to automotive industry. In 2015 10th Asian Control Conference (ASCC), IEEE, 2015, pp. 1–5.
- [18] Ghosh, R.K., Banerjee, A., Aich, P., Basu, D. and Ghosh, U., 2022. Intelligent IoT for Automotive Industry 4.0: Challenges, Opportunities, and Future Trends. *Intelligent Internet of Things for Healthcare and Industry*, pp.327-352.
- [19] Kozma, J., 1998. Intelligent agents. *IEEE Potentials*, 17(2), pp.16-19.
- [20] Ghosh, R.K., Banerjee, A., Aich, P., Basu, D. and Ghosh, U., 2022. Intelligent IoT for Automotive Industry 4.0: Challenges, Opportunities, and Future Trends. *Intelligent Internet of Things for Healthcare and Industry*, pp.327-352.
- [21] Atif, Y., Ding, J. and Jeusfeld, M.A., 2016. Internet of things approach to cloud-based smart car parking. *Procedia Computer Science*, 98, pp.193-198.
- [22] Kshetrimayum, R.S., 2009. An introduction to UWB communication systems. *Ieee Potentials*, 28(2), pp.9-13.
- [23] Patel, K.R. and Kulkarni, R., 2014. Ultra-wideband (UWB) wireless system. *International Journal of Application or Innovation in Engineering & Management*, pp.1-15.
- [24] Balakrishnan, D., Nayak, A., Dhar, P. and Kaul, S., 2009, June. Efficient geo-tracking and adaptive routing of mobile assets. In 2009 11th IEEE International Conference on High Performance Computing and Communications (pp. 289-296). IEEE.
- [25] Liu, Z., Zhang, A. and Li, S., 2013, July. Vehicle anti-theft tracking system based on Internet of things. In *Proceedings of 2013 IEEE International Conference on Vehicular Electronics and Safety* (pp. 48-52). IEEE.
- [26] Rahiman, W. and Zainal, Z., 2013, June. An overview of development GPS navigation for autonomous car. In 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA) (pp. 1112-1118). IEEE.
- [27] Rico, J., Sancho, J., Cendon, B. and Camus, M., 2013, March. Parking easier by using context information of a smart city: Enabling fast search and management of parking resources. In 2013 27th international conference on advanced information networking and applications workshops (pp. 1380-1385). IEEE.
- [28] Thomas, M.O. and Rad, B.B., 2017. Reliability evaluation metrics for internet of things, car tracking system: a review. *Int. J. Inf. Technol. Comput. Sci.(IJITCS)*, 9(2), pp.1-10.

# The role of augmented and virtual reality in improving consumer marketing experiences

Mihalj Bakator<sup>1</sup>, Luka Đorđević<sup>1</sup>, Borivoj Novaković<sup>1</sup>, Mića Đurđev<sup>1</sup> and Stefan Ugrinov<sup>1</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

*mihalj.bakator@tfzr.rs; luka.djordjevic@tfzr.rs; borivoj.novakovic@gmail.com; mica.djurdjev@tfzr.rs; stefan.ugrinov@tfzr.rs*

## Abstract:

This paper analyzes the transformative potential of Augmented Reality (AR) and Virtual Reality (VR) technologies in enhancing customer marketing experiences. AR and VR offer innovative, interactive, and personalized experiences that traditional marketing methods lack. Despite their promise, challenges such as high costs, technical limitations, and privacy concerns hinder widespread adoption. Governments can facilitate this adoption through investment in digital infrastructure, financial incentives, and robust data protection regulations. Enterprises should focus on creating high-quality AR and VR content, training employees, and ensuring data security. Collaboration between businesses and technology developers is important for optimizing these technologies. Consumers also play a role by embracing AR and VR, providing feedback, and maintaining privacy awareness. A collective effort among governments, enterprises, and individuals is essential for overcoming barriers and maximizing the benefits of AR and VR in marketing. This paper concludes that AR and VR have the potential to significantly improve customer experiences and business outcomes, heralding a new era of innovative and immersive marketing strategies.

## Keywords:

Augmented reality, virtual reality, consumer experience, marketing

## 1. Introduction

Augmented Reality (AR) and Virtual Reality (VR) technologies have increasingly become significant in the realms of marketing and user experience, revolutionizing the way businesses engage with their customers. These immersive technologies offer innovative avenues for creating interactive and memorable experiences that traditional marketing strategies often fail to deliver. By blending the physical and digital worlds, AR and VR provide marketers with powerful tools to captivate and inform consumers in novel ways [1]. AR technology overlays digital information onto the physical world, enriching the real-world environment with virtual elements. This capability enables marketers to deliver improved product information, interactive advertisements, and personalized shopping experiences directly through smartphones or AR glasses. For instance, AR can allow customers to visualize how furniture would look in their homes, try on virtual clothes, or receive detailed product specifications by simply pointing their devices at an item. This level of interactivity not only engages consumers but also aids them in making informed purchasing decisions, thereby improving overall customer satisfaction [2]. Conversely, VR creates entirely digital environments, offering users a sense of presence in a completely virtual space. This immersive experience is particularly valuable for product demonstrations, virtual tours, and experiential marketing campaigns. By allowing consumers to explore products or services in a fully controlled virtual setting, VR can capture their attention and provide a comprehensive understanding of what is being offered. For example, automotive companies use VR to let potential customers virtually test-drive cars, while real estate firms offer virtual tours of properties, enabling clients to experience homes from the comfort of their own spaces [3]. The integration of AR and VR into marketing strategies has shown immense potential in enhancing customer engagement, satisfaction, and ultimately, sales. By leveraging these technologies, companies can deliver unique and personalized experiences that resonate with consumers on a deeper level. Furthermore, the data collected from AR and VR interactions provides valuable insights into consumer preferences and



behaviors. This information can be instrumental in developing more targeted and effective marketing campaigns, allowing businesses to tailor their efforts to meet the specific needs and desires of their audience [4].

As AR and VR technologies continue to advance, their role in shaping the future of marketing and user experience is set to expand. These technologies are expected to offer even more sophisticated and impactful ways to connect with audiences. From creating virtual showrooms to enabling interactive brand storytelling, the possibilities are vast and varied. The continued evolution of these technologies promises to redefine consumer interactions, offering unprecedented opportunities for engagement and growth in the marketing sector [5].

The paper consists of five main sections. The first section introduces the subject. Next, augmented and virtual reality are addressed. The third section discusses improving consumer marketing experience. After that, suggestions and guidelines are noted. Finally, conclusions are drawn and guides for future research are noted.

## 2. Augmented and virtual reality

Augmented Reality (AR) and Virtual Reality (VR) technologies have garnered significant attention for their potential to revolutionize various industries, including marketing, healthcare, education, and entertainment. These technologies create immersive experiences that blend or replace the real world with digital environments. Despite their promise, several challenges hinder their widespread application [6]. One of the main challenges in the application of AR and VR is the high cost of development and implementation. Creating high-quality AR and VR experiences requires substantial investment in hardware, software, and skilled professionals. Devices such as VR headsets and AR glasses can be expensive, making them less accessible to the general public. Additionally, the development of detailed and interactive virtual environments demands significant time and resources, often limiting these technologies to larger companies with substantial budgets [7].

Another challenge is the technical limitations and the need for powerful computing capabilities. Both AR and VR applications require robust processing power to render complex graphics and ensure smooth, real-time interactions. This necessity can lead to performance issues, such as latency and motion sickness, which can detract from the user experience. Moreover, ensuring compatibility across various devices and platforms adds another layer of complexity, as developers must optimize their applications for different hardware specifications and operating systems [8]. User acceptance and adaptation also pose significant hurdles. Many potential users may be unfamiliar with AR and VR technologies, leading to resistance or hesitation in adopting them. There is also a learning curve associated with using these technologies effectively, which can be a barrier for widespread use. Privacy concerns and data security issues further complicate the landscape, as AR and VR applications often collect and process large amounts of personal data, raising questions about user privacy and the potential for misuse [9]. Despite these challenges, AR and VR offer numerous advantages that make them attractive for various applications. In marketing, these technologies enable companies to create engaging and interactive experiences that can captivate consumers and provide a deeper understanding of products or services. For example, AR can allow customers to visualize how furniture would look in their home, while VR can offer virtual test drives for cars, enhancing the decision-making process [10].

In education, AR and VR can transform learning by providing immersive environments that facilitate hands-on experiences and interactive lessons. Students can explore historical sites, conduct virtual lab experiments, or engage in simulations that would be impossible in a traditional classroom setting. This level of engagement can improve comprehension and retention of information, making learning more effective and enjoyable [11].

In healthcare, AR and VR technologies are being used for training medical professionals, planning surgeries, and even treating patients. These technologies allow for detailed visualization of anatomical structures, enabling more precise and effective medical interventions. VR therapy is also being explored for treating mental health conditions, providing a controlled and safe environment for exposure therapy and other treatments [11].

### 3. Improving consumer marketing experiences

Augmented Reality (AR) and Virtual Reality (VR) technologies are profoundly transforming consumer marketing by creating more engaging, interactive, and personalized experiences. These immersive technologies offer unique opportunities for brands to connect with consumers in ways that traditional marketing methods cannot match. By integrating AR and VR into marketing strategies, companies can significantly improve the way consumers interact with products, leading to improved satisfaction, stronger brand loyalty, and increased sales [12]. AR improves the physical environment by overlaying digital elements, allowing consumers to interact with products in an augmented space. This technology is being utilized in various industries to provide richer shopping experiences. For instance, furniture retailers use AR applications to enable customers to visualize how a piece of furniture would look in their homes. By simply using a smartphone or tablet, users can place virtual furniture in their actual living spaces, adjusting the size, color, and position to see how it fits with their existing decor. This not only helps customers make more informed purchasing decisions but also reduces the likelihood of returns, as they have a clearer understanding of how the product will look and feel in their environment [13]. AR is making significant strides in the beauty and fashion industries. Consumers can use AR applications to virtually try on makeup, glasses, clothing, and accessories. This capability allows them to see how different products look on them without physically trying them on, making online shopping more interactive and reducing the uncertainty often associated with buying personal items online. For example, beauty brands are leveraging AR to create virtual mirrors that let users experiment with various makeup shades and styles, instantly seeing the results on their faces. Similarly, eyewear companies offer AR experiences where customers can try on different frames to find the perfect fit. This not only improves the convenience and enjoyment of the shopping experience but also increases customer loyalty and repeat purchases by providing a personalized and engaging shopping journey [14].

On the other hand, VR creates fully immersive digital environments where consumers can engage with products or services in a controlled, virtual space. This is particularly useful for high-involvement purchases such as automobiles or real estate. Car manufacturers are utilizing VR to provide virtual test drives, allowing potential buyers to experience the look, feel, and performance of a vehicle without needing to visit a dealership. This immersive experience can include navigating through different terrains, exploring the interior features, and even customizing the car's specifications, offering a comprehensive understanding of the product. Real estate companies offer virtual tours of properties, enabling clients to explore homes and commercial spaces from the comfort of their own devices. Prospective buyers can navigate through rooms, examine details, and get a true sense of the property's layout and ambiance. These immersive experiences provide a deeper level of engagement and can significantly influence purchasing decisions by allowing consumers to experience products in a realistic and detailed manner [15].

Additionally, VR is being leveraged for experiential marketing campaigns that create memorable brand experiences. Companies can design virtual events, product launches, or branded environments that consumers can explore. These experiences can be tailored to provide personalized interactions based on user preferences and behaviors, creating a strong emotional connection with the brand. For example, a fashion brand could create a virtual runway show that allows users to view new collections from different angles, interact with the models, and even purchase items directly from the virtual event. Similarly, an automotive brand could host a virtual car show where attendees can explore the latest models, watch performance demonstrations, and engage with interactive content. Such immersive marketing efforts can generate buzz and word-of-mouth promotion, further extending the reach and impact of marketing campaigns [3].

AR and VR technologies offer valuable data collection opportunities. By tracking how consumers interact with virtual products and environments, companies can gain deep insights into consumer preferences, behaviors, and pain points. This data can be used to refine marketing strategies, develop more targeted campaigns, and improve product designs to better meet consumer needs. For instance, analyzing which virtual products are tried on most frequently can inform inventory decisions and highlight popular trends. Retailers can also use heat maps and interaction metrics to understand which

areas of a virtual store attract the most attention, optimizing store layouts and product placements accordingly [16].

The integration of AR and VR into consumer marketing also supports sustainability efforts. Virtual showrooms and product demonstrations reduce the need for physical samples, printed materials, and travel, lowering the environmental footprint of marketing activities. This can appeal to environmentally conscious consumers and improve the brand's reputation for sustainability. For example, automotive brands can conduct virtual car launches, reducing the need for transporting vehicles and setting up physical event spaces. Fashion brands can use virtual fashion shows to showcase new collections, minimizing the waste associated with traditional runway events [17].

AR and VR can improve customer support and after-sales service. For instance, AR can provide step-by-step visual guides for product assembly or troubleshooting, allowing customers to resolve issues quickly and efficiently. VR can offer immersive training modules for customer service representatives, helping them to better understand the products and deliver superior support. These applications not only improve the customer experience but also reduce the burden on customer support teams, leading to more efficient and effective service [18]. Furthermore, a model for improving consumer marketing experiences through augmented and virtual reality is developed. The model is presented on Figure 1.

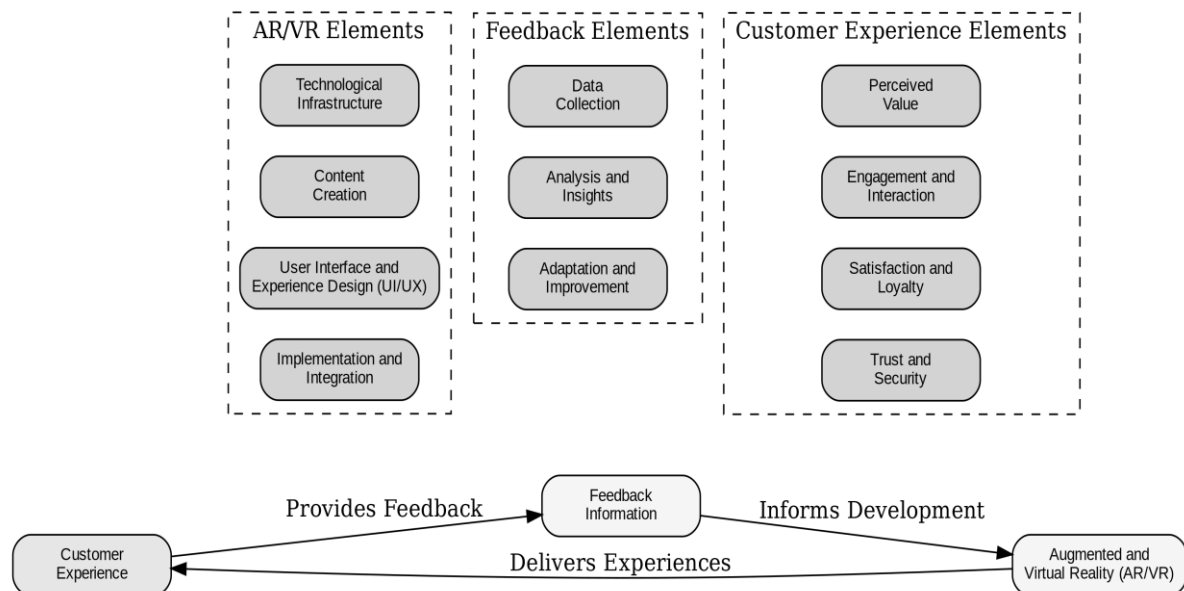


Figure 1: Model for improving consumer marketing experiences through augmented and virtual reality

AR/VR Elements serve as the foundational technological and content components required for delivering AR and VR experiences. This includes technological infrastructure, content creation, user interface and experience design (UI/UX), and the implementation and integration of these elements. These components are essential for creating immersive and interactive experiences that engage customers.

Feedback Elements are responsible for collecting, analyzing, and utilizing data to improve AR/VR experiences. Data collection involves gathering information on how users interact with the AR/VR environment. Analysis and insights refer to processing this data to understand user behavior and preferences. Adaptation and improvement involve modifying the AR/VR elements based on these insights to improve user satisfaction and engagement.

Customer Experience Elements reflect the impact of AR/VR experiences on users. These include perceived value, which indicates how much users feel they gain from the experience, and engagement and interaction, which measure the level of user involvement. Satisfaction and loyalty are outcomes of positive experiences, while trust and security ensure that users feel safe and confident in using the technology.

The relationship between these elements is cyclical. AR/VR experiences are delivered to customers, who then provide feedback through their interactions. This feedback is collected and analyzed, providing valuable information that informs the development and refinement of AR/VR elements. This continuous loop ensures that the experiences are continually optimized, leading to improved customer satisfaction and loyalty over time.

The diagram illustrates the relationship between AR/VR elements, feedback mechanisms, and customer experience components, creating a cycle that continuously improves user satisfaction. Technological infrastructure encompasses the hardware and software needed to support AR/VR environments, ensuring high-quality visuals, responsive interactions, and stable connections. Content creation involves developing engaging and interactive materials like 3D models and virtual environments, which are important for maintaining user engagement and immersion.

User interface and experience design (UI/UX) focuses on creating intuitive interfaces, allowing users to interact easily with the AR/VR environment. This is essential for reducing frustration and maximizing engagement. Implementation and integration bring together various technological and content components to form a cohesive AR/VR experience. This process involves testing and optimizing the system to ensure smooth performance across different devices.

Feedback elements play an important role in refining the AR/VR experience. Data collection gathers information on user interactions and behaviors, providing a foundation for analysis. Analysis and insights then interpret this data to identify patterns and areas for improvement. Adaptation and improvement follow, where AR/VR elements are adjusted based on these insights to better align with user expectations, ensuring the experience remains relevant and enjoyable.

Customer experience elements measure the impact of AR/VR on users. Perceived value reflects how much users benefit from the experience, influenced by content quality and usability. Engagement and interaction indicate the level of user involvement, with high engagement leading to memorable experiences. Satisfaction and loyalty result from positive experiences, encouraging users to return. Trust and security are important for user confidence, ensuring safe and reliable performance, which is particularly important for applications handling sensitive data. Together, these elements form a continuous feedback loop that drives the development and improvement of AR/VR experiences.

## 4. Suggestions and guidelines

Next, in this section guidelines and suggestions for improving consumer marketing experiences through augmented and virtual reality are noted:

- Invest in digital infrastructure such as high-speed internet and 5G networks to support AR and VR applications.
- Provide grants and tax incentives for companies developing AR and VR technologies.
- Establish clear regulations and guidelines to protect user data in AR and VR applications.
- Support ongoing research and development through funding and collaboration with academic institutions.
- Integrate AR and VR into marketing strategies to improve customer engagement and satisfaction.
- Invest in creating high-quality AR and VR content, such as virtual product demonstrations and interactive advertisements.
- Train employees to effectively use AR and VR technologies in customer interactions.
- Collaborate with AR and VR developers to create seamless and user-friendly applications.
- Prioritize data security by implementing robust cybersecurity measures and being transparent about data collection practices.
- Make AR and VR technologies more affordable through subsidies or financing options for devices.
- Regularly update AR and VR applications based on user feedback and technological advancements.
- Use AR for real-time, visual support in customer service, and VR for immersive training modules for customer representatives.
- Embrace AR and VR technologies for more interactive and personalized shopping experiences.

- Provide feedback on AR and VR experiences to help companies refine their offerings.
- Participate in beta testing and provide constructive feedback to developers.
- Governments and enterprises should work together to improve accessibility of AR and VR technologies.
- Support innovation through industry forums and consortia to stay updated on advancements and best practices.
- Ensure continuous improvement by investing in infrastructure, supporting innovation, prioritizing privacy, and ensuring accessibility.

## 5. Conclusions

In conclusion, the integration of Augmented Reality (AR) and Virtual Reality (VR) technologies presents a transformative opportunity for enhancing customer marketing experiences. These technologies offer unprecedented levels of interactivity, personalization, and immersion, enabling businesses to connect with consumers in innovative ways. Despite the promising potential, several challenges, such as high costs, technical limitations, and privacy concerns, must be addressed to facilitate widespread adoption. Governments play a important role by investing in digital infrastructure, providing financial incentives for AR and VR development, and establishing robust data protection regulations. Enterprises must focus on creating high-quality, user-friendly AR and VR content, training employees, and ensuring data security to build consumer trust and satisfaction. Collaboration between businesses and technology developers is essential for optimizing AR and VR applications and staying abreast of technological advancements.

For future studies, the application effects of AR and VR in enterprises could be addressed. This would provide additional valuable insight into this subject.

## Acknowledgement

This paper has been supported by the Provincial Secretariat for Higher Education and Scientific Research of the Autonomous Province of Vojvodina, number: 142-451-2963/2023-01.

## References:

- [1] S. T. Dlamini, C. Botha, The use of augmented and virtual reality to improve the customer experience, *African Journal of Science, Technology, Innovation and Development* 16 (2024) 399–409. doi:10.1080/20421338.2024.2326791.
- [2] J.-Y. Ho, G. Ju, S. Hong, J. An, C. C. Lee, Factors influencing customer satisfaction with AR shopping assistant applications in e-commerce: An empirical analysis utilizing text-mining techniques, *Aslib Journal of Information Management* (2023). doi:10.1108/AJIM-03-2023-0089.
- [3] A. de Regt, K. Plangger, S. J. Barnes, Virtual reality marketing and customer advocacy: Transforming experiences from story-telling to story-doing, *Journal of Business Research* 136 (2021) 513–522. doi:10.1016/j.jbusres.2021.08.004.
- [4] W. D. Hoyer, M. Kroschke, B. Schmitt, K. Kraume, V. Shankar, Transforming the customer experience through new technologies, *Journal of Interactive Marketing* 51 (2020) 57–71. doi:10.1016/j.intmar.2020.04.0.
- [5] D. Adams, Virtual retail in the metaverse: Customer behavior analytics, extended reality technologies, and immersive visualization systems, *Linguistic and Philosophical Investigations* 21 (2022) 73–88.
- [6] M. C. tom Dieck, T. H. Jung, S. M. Loureiro, Augmented reality and virtual reality: New trends in immersive technology, *Springer Nature* (2021).
- [7] C. Boletsis, A. Karahasanovic, Immersive technologies in retail: Practices of augmented and virtual reality, *Proceedings of the 4th International Conference on Computer-Human Interaction Research and Applications* (2020). doi:10.1007/11250/3084393.
- [8] C. Naji, A. A. Karakhan, Technologies for safety and health management in construction: Current use, implementation benefits and limitations, and adoption barriers, *Journal of Building Engineering* 29 (2020) 101212. doi:10.1016/j.jobbe.2020.101212.

- [9] M.-C. Voicu, N. Sîrghi, D. M.-M. Toth, Consumers' experience and satisfaction using augmented reality apps in E-shopping: New empirical evidence, *Applied Sciences* 13 (2023) 9596. doi:10.3390/app13179596.
- [10] A. Jessen, T. Hilken, M. Chylinski, D. Mahr, J. Heller, D. I. Keeling, K. de Ruyter, The playground effect: How augmented reality drives creative customer engagement, *Journal of Business Research* 116 (2020) 85–98. doi:10.1016/j.jbusres.2020.05.002.
- [11] Z. Qu, C. W. Lau, S. J. Simoff, P. J. Kennedy, Q. V. Nguyen, D. R. Catchpoole, Review of innovative immersive technologies for healthcare applications, *Innovations in Digital Health, Diagnostics, and Biomarkers* 2 (2022) 27–39. doi:10.36401/IDDB-21-04.
- [12] K. Khanal, Crafting sustainable brand narratives through immersive technologies: The role of virtual reality (VR) and augmented reality (AR), in: *Compelling Storytelling Narratives for Sustainable Branding*, IGI Global (2024), pp. 134–144.
- [13] L. Xue, Designing effective augmented reality platforms to improve the consumer shopping experiences [PhD Thesis], Loughborough University (2022).
- [14] M.-C. Voicu, N. Sîrghi, D. M.-M. Toth, Consumers' experience and satisfaction using augmented reality apps in E-shopping: New empirical evidence, *Applied Sciences* 13 (2023) 9596. doi:10.3390/app13179596.
- [15] G. Branca, V. Marino, R. Resciniti, How do consumers evaluate products in virtual reality? A literature review for a research agenda, *Spanish Journal of Marketing-ESIC* 28 (2024) 356–380. doi:10.1108/SJME-07-2022-0153.
- [16] D. A. Wieland, B. S. Ivens, E. Kutschma, P. A. Rauschnabel, Augmented and virtual reality in managing B2B customer experiences, *Industrial Marketing Management* 119 (2024) 193–205. doi:10.1016/j.indmarman.2024.04.007.
- [17] G. H. Popescu, K. Valaskova, J. Horak, Augmented reality shopping experiences, retail business analytics, and machine vision algorithms in the virtual economy of the metaverse, *Journal of Self-Governance and Management Economics* 10 (2022) 67–81.
- [18] B. Romano, S. Sands, J. I. Pallant, Augmented reality and the customer journey: An exploratory study, *Australasian Marketing Journal* 29 (2021) 354–363. doi:10.1016/j.ausmj.2020.06.010.

# Investigating IT Experts' Attitudes Toward the Frequency of Agile Meetings: A Thematic Analysis

Maja Gaborov<sup>1</sup>, Zeljko Stojanov<sup>1</sup>, Srdjan Popov<sup>2</sup>, Dragana Kovac<sup>1</sup>, Igor Vecstejn<sup>1</sup>, Gordana Jotanovic<sup>3</sup>, Goran Jausevac<sup>3</sup>

<sup>1</sup> University of Novi Sad, Tehnical Faculty "Mihajlo Pupin", DjureDjakovića bb, Zrenjanin, Serbia

<sup>2</sup> University of Novi Sad, Faculty of Technical Sciences, TrgDositejaObradovica 6, Novi Sad, Serbia

<sup>3</sup> University of East Sarajevo, Faculty of Transport and Traffic Engineering, Dobo, Bosnia and Herzegovina

maja.gaborov@tfzr.rs; zeljko.stojanov@uns.ac.rs; srdjanpopov@uns.ac.rs;  
dragana.milosavljev@tfzr.rs; igor.vecstejn@tfzr.rs; gordana.jotanovic@sf.ues.rs.ba;  
goran.jausevac@sf.ues.rs.ba

## Abstract:

This study employs a qualitative thematic analysis methodology to scrutinize the impact of meeting frequency on employees' attitudes toward meetings. Semi-structured interviews were conducted with IT experts, exploring their experiences, thoughts, and feelings regarding meeting frequency. Through detailed thematic analysis, key themes emerged, providing insight into how meeting frequency influences employee engagement, productivity, and satisfaction. The results reveal a prevailing negative sentiment among participants towards meeting frequency. The findings are presented as themes that will be useful for further research.

## Keywords:

IT experts, Attitudes, Agile, Meeting Frequency, Thematic analysis

## 1. Introduction

Agile methodologies emphasize flexibility, collaboration, and continuous improvement [1]. In agile frameworks, iterative processes and frequent collaboration emphasize the importance of effective communication and coordination. Agile meetings are essential for synchronizing activities, discussing progress, and resolving obstacles [2]. These focused meetings encourage communication, transparency, and collaboration among team members, facilitating quick decision-making and problem-solving. While agile methodologies promote the value of regular meetings for team effectiveness, the frequency of these meetings varies across organizations [3]. Understanding how meeting frequency influences employee attitudes is critical for optimizing organizational productivity and engagement. Frequent meetings may enhance coordination and alignment, leading to increased productivity. However, excessive meetings can result in disengagement, and reduced productivity [4]. In this context, qualitative research emerges as a valuable approach for exploring the relationship between meeting frequency and attitudes. Semi-structured interviews offer flexibility and depth in gathering insights directly from participants, ensuring a comprehensive understanding of their perspectives and experiences [5, 6, 7]. This approach allows for a thorough exploration of how meeting frequency impacts employee attitudes toward meetings.

In this study, semi-structured interviews were conducted with IT experts to investigate how the frequency of meetings affects attitudes towards meetings within agile methodologies. By examining the experiences of IT experts, the aim is to discover the links between the frequency of encounters and the attitudes of employees. It will be presented based on a thematic analysis based on which key themes are presented.

## 2. Related work

Research on meetings has long recognized the importance of meeting frequency in shaping employee perceptions and behaviors. For example, in one study [8] the author found that frequent meetings were associated with increased negative attitudes and decreased job satisfaction. Similarly, [3] observed that high meeting frequency negatively affected team dynamics and performance. However, this study [9] also highlighted the role of meeting quality in moderating the relationship between frequency and attitudes in earlier work, an inductive thematic analysis was performed to determine the themes that represent problems in Agile meetings. In the mentioned work, it was possible to see how themes are determined based on qualitative data analysis. Furthermore, in one of the studies [10], the author investigated the relationship between meeting frequency and organizational communication climate through interviews with IT experts. Their study revealed that organizations with a culture of frequent meetings often experienced communication overload and diminished employee satisfaction. Participants highlighted the need for clear communication channels and streamlined meeting processes to mitigate the negative effects of meeting frequency on attitudes and performance.

## 3. Methodology

This research aims to investigate the impact of meeting frequency on the attitudes of employed IT experts through thematic analysis. The primary goal of this research is to investigate how the frequency of agile meetings affects the attitudes of IT experts towards their work, and the main contribution is a thematic analysis that reveals the key challenges associated with frequent meetings.

### 3.1 Sampling of participants

Seven IT experts, with experience ranging from 3 to 11 years, were interviewed to express their views on meeting frequency, covering software developer, mid- and senior-level roles. The IT experts are from medium and larger companies that apply agile methodologies, mainly Scrum. Sample size considerations in qualitative research, such as data saturation, ensure adequacy [11, 12]. In this study, saturation signaled sufficient participant involvement, enhancing the rigor and validity of the study [13, 14]. The anonymity of the participants promoted an open dialogue, aligned with the ethical principles of informed consent and confidentiality [15, 16].

### 3.2 Data collection

Data collection for this study involves individual, in-depth, semi-structured interviews, recorded and transcribed. Participants' views were transcribed in a Word document [17]. We chose interviews for their flexibility in addressing ambiguities and prompting elaboration [18]. Transcription is integral for subsequent data analysis [19], especially in semi-structured interviews requiring follow-up inquiries [20]. Interviews were conducted in person or via Zoom, offering convenience and flexibility [21]. Participants were fully informed and provided informed consent to ensure voluntary participation and confidentiality.

### 3.3 Thematic data analysis

A qualitative analysis is conducted to delve into participants' perspectives. Thematic analysis played a key role in our research. This systematic approach, as presented in the study [22], made it possible to examine the interview data in detail. Analysis involved coding the data, deriving participant attitudes, and synthesizing codes into themes, employing a structured six-stage methodology [23]. During the thematic analysis, we came to the main topic, "The impact of the



frequency of meetings on the attitudes of IT experts". The main topic is broken down into several subthemes. Thematic analysis of the interview data revealed several sub-themes:

- Burden of Meetings
- Effect on Productivity
- Inconvenience and Time Management
- Boredom and Disengagement
- Lack of Continuity and Impact on Team Dynamics.

Figure number 1 presents the main theme and the sub-themes that emerged during the thematic analysis.

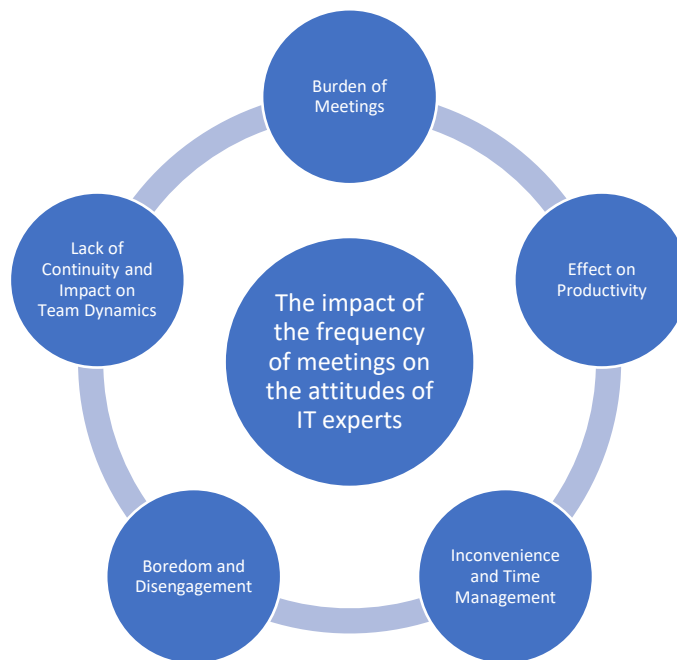


Figure 1: Presentation of the main theme and sub-themes in the analysis of problems in agile meetings

Based on the codes, sub-themes were determined from which one of the main mentioned themes emerged. The figure 2 shows the main theme, sub-theme, and the codes from which the sub-themes and the theme were derived. Subthemes are represented by light blue squares, and codes by a white circles. Each sub-theme emerged from the codes. Each sub-theme is connected with an arrow to specific codes since the sub-theme was created based on similar codes.

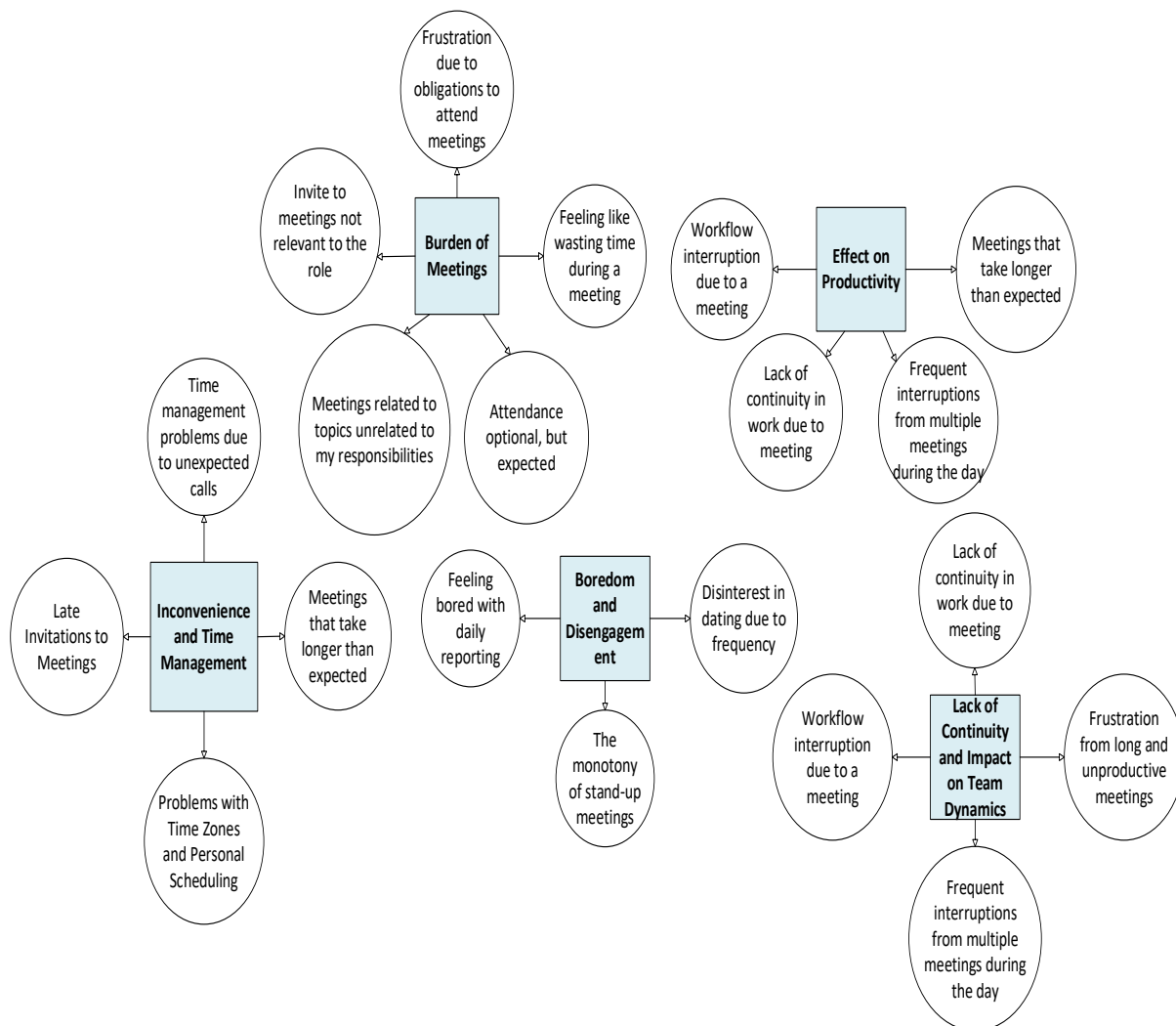


Figure 2: Thematic analysis of the impact of the frequency of meetings on the attitudes of IT expert

Participants express frustration regarding the burden of attending numerous meetings, which they perceive as detracting from their regular commitments and work tasks. They describe instances where they receive invitations to meetings that they deem unnecessary or unrelated to their roles, leading to discomfort and a sense of obligation to attend. Participants highlight the negative impact of frequent meetings on productivity and workflow. They describe how interruptions and a lack of continuity due to multiple meetings throughout the day hinder their ability to focus and complete tasks efficiently. The constant shift between meetings leaves them feeling unproductive and overwhelmed. Participants discuss the inconvenience caused by late or unexpected meeting invitations, which disrupt their planned work schedules and personal time. They cite examples of managers scheduling meetings without consideration for time zones or providing short notice, resulting in difficulties in time management and work-life balance. Participants express feelings of boredom and disengagement during meetings, particularly daily stand-up meetings. They perceive these routine meetings as repetitive and unnecessary, contributing to a negative attitude towards attending them regularly. Participants express frustration with the lack of continuity and coherence in their work due to frequent interruptions from meetings.

### 3.3.1 Discussion of Themes

The thematic analysis provides valuable insights into the impact of meeting frequency on employee attitudes. While meetings are essential for collaboration and communication, excessive

meeting frequency can have detrimental effects on employee well-being and productivity. All participants said that the frequency of the meeting harmed their attitudes towards the meeting.

The participant stated: *'People don't want to come to meetings because they lose the time they need to deal with their regular responsibilities'*. This highlights that people often feel they are wasting time in meetings because they have other work activities during the day.

The next participant is discussing various challenges and frustrations they face regarding meeting invitations and their experiences with attending meetings. They mention situations where they receive invitations to meetings that they feel are unnecessary or irrelevant to their roles within the company. Additionally, they describe instances where discussions in meetings pertain to topics unrelated to their responsibilities, causing discomfort and a sense of obligation to attend. The participant also addresses issues such as the accumulation of meetings, disruptions caused by late invitations, and the negative impact of meetings extending beyond expected durations. Furthermore, they express frustration with meetings discussing topics that may not directly involve them and highlight the inconvenience of scheduling conflicts and time zone differences. Overall, the participant's comments reflect their negative attitude towards meetings, which is influenced by various factors such as relevance, frequency, and duration. Participant stated: *'It may happen that you get invited to meetings that you don't think you should attend. With us, it happens that someone talks about a completely different branch of the company that has nothing to do with yours, where the manager emphasizes that it is not mandatory for you to attend, but in principle he wants you to be present. We also have meetings that are once a month, but when it all adds up, there are a lot of them. I believe that HR should call us when someone asks for help, when there is a problem. When the meetings pile up, I don't see the point of being in those meetings because they are frequent and take some time away from other activities, they don't last just 15 minutes, but up to an hour. It can also happen that I am called late for a meeting, where the manager makes an appointment for today. It has to do with the time zone in my case because the manager makes an appointment at ten in the evening for the appointment when he wakes up, I try to plan to go to work at 9 and be there by 5, but at 11 in the morning I see that he has an appointment for 7 in the evening and therefore I will be at work for ten hours. Therefore, when there are many meetings, a negative attitude develops.'*

One of the participants expressed feeling bored by the daily obligation to discuss their daily activities, regardless of the meeting's duration, suggesting a sense of monotony and disinterest in routine reporting. The participant stated the following: *'It doesn't matter if today lasts a long time, ten minutes, or an hour, but when I know that I should talk about what happened to me during the day every day at nine o'clock, I'm bored.'*

The next participant criticizes the negative impact of frequent interruptions from multiple meetings on productivity and workflow, highlighting the frustration of losing continuity and feeling unproductive due to the fragmented nature of their workday. *'The frequency of meetings is worse than when there are long meetings because you are constantly interrupted by someone and you do not have continuity in your work. I feel bad, and I also see people from my team who do not like to have frequent meetings at all, especially if they are several times a day and if they are not continuous. Literally, you rest a little during the break and then you have a meeting again. Time goes by like that, and you haven't done anything.'*

#### 4. Research implications

Researchers reading this paper should gain insights into the significance of meeting frequency on employee attitudes. They should learn about the complexities involved in managing meeting schedules and how these factors can impact employee engagement, productivity, and job satisfaction. Additionally, researchers should understand the importance of employing qualitative methodologies, such as thematic analysis, to gain deeper insights into participants' experiences. This paper underscores the need for future research to explore practical strategies for optimizing meeting practices.

Practitioners should carefully consider the necessity and purpose of each meeting, aiming to minimize unnecessary meetings and prioritize those that are essential for achieving organizational goals. Strategic meeting planning can help alleviate meeting overload and enhance employee

satisfaction. Organizations should invest in training for meeting facilitators and participants to improve meeting effectiveness and engagement. Providing resources and guidance on meeting best practices can help ensure that meetings are well-structured, purposeful, and conducive to collaboration. Practitioners should recognize that employees' needs and preferences regarding meeting frequency may vary. Offering flexibility in meeting schedules and formats, such as allowing for virtual participation or asynchronous communication, can accommodate diverse work styles and preferences. Cultivating a culture of respect for employees' time and contributions is essential for fostering positive attitudes towards meetings. Organizations should encourage open communication and feedback channels, allowing employees to voice their concerns and suggestions regarding meeting practices.

## 5. Trustworthiness

Trustworthiness encompasses the credibility, reliability, and dependability of research findings and interpretations, instilling confidence in the accuracy, integrity, and transparency of the research process and outcomes [22]. In our study examining the impact of meeting frequency on attitudes towards meetings, we took several measures to ensure trustworthiness. Acknowledging the limited sample size in our study, we ensured meticulous definition of methodological questions and objectives, aligning with the notion that a small sample size can suffice under such circumstances [24]. To establish credibility, we employed various strategies. Firstly, we conducted semi-structured interviews with open-ended questions, allowing participants to freely articulate their thoughts and experiences. Additionally, we fostered rapport with participants, creating an environment of trust and open communication. The active engagement of all authors in discussions and findings development ensured consensus, enhancing the confidentiality of the study findings. We believe that the insights gleaned from our study hold relevance beyond the specific context, potentially benefiting other IT companies. To ensure reliability, we adhered to a systematic approach to data collection and analysis. Thematic analysis followed established guidelines, with the entire process meticulously described. Interviews were transcribed verbatim, and multiple researchers independently analyzed the data to identify themes. While participants expressed a perception that frequent meetings disrupt their workflow, leading to feelings of decreased productivity, the study did not directly measure productivity. This presents an opportunity for future research to quantitatively evaluate meeting frequency's impact on productivity metrics. Any discrepancies were resolved through thorough discussion among the authors. These measures contribute to the reliability and trustworthiness of our study findings.

## 6. Conclusion

In conclusion, this study clarifies the relationship between meeting frequency and employee attitudes toward meetings. Through understanding employee experiences, organizations can refine meeting practices to enhance engagement, productivity, and satisfaction in the workplace. This paper presents an analysis of interviews conducted with IT industry experts to explore the impact of meeting frequency on participants' views. Frequent meetings can consume a significant amount of employees' time, leaving less time for them to focus on their regular tasks. This can lead to frustration and a negative attitude toward meetings. When meetings are held frequently throughout the day, it can disrupt employees' workflow and concentration. Constant interruptions can hinder productivity and lead to resentment toward meetings. If meetings are held too often or are not perceived as adding value, employees may view them as a waste of time. This can contribute to a negative attitude toward meetings and a lack of motivation to participate. Excessive meeting frequency can contribute to feelings of burnout among employees, as they may feel overwhelmed by the constant demand on their time and attention. This can lead to resentment toward meetings and a desire to avoid them whenever possible. Frequent meetings can contribute to communication overload, where employees feel bombarded with information and updates. This can lead to feelings of frustration and disengagement, particularly if the information being shared is not relevant to their work. Quantitative analysis of the survey data will be conducted in the future to increase confidence in the findings. More participants

will be involved. We also plan to include additional participants from different organizational roles, such as team leaders, HR staff, and senior management, to gain an understanding of how meeting frequency affects different stakeholders.

### References:

- [1] K. Beck, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, et al., "Manifesto for agile software development," *Agile Alliance*, 2001.
- [2] K. Schwaber, D. Sutherland, *The Scrum Guide: The definitive guide to Scrum: The rules of the game*, Scrum.org, 2017.
- [3] C. Jones, L. Wang, "The role of meeting frequency in agile software development teams," *Journal of Agile Development*, vol. 8, no. 2, pp. 34–47, 2019.
- [4] S. G. Rogelberg, D. J. Leach, P. B. Warr, J. L. Burnfield, "Not another meeting! Are meeting distractions affecting your bottom line?" *Business Horizons*, vol. 49, no. 1, pp. 39–46, 2006.
- [5] A. B. Hamilton, E. P. Finley, "Qualitative methods in implementation research: An introduction," *Psychiatry Research*, vol. 280, p. 1125, 2019.
- [6] J. M. Corbin, A. L. Strauss, *Basics of Qualitative Research*, 4th ed., Sage Publications, California, 2015.
- [7] R. Kumar, *Research Methodology: A Step-By-Step Guide for Beginners*, 3rd ed., SAGE Publications, 2011.
- [8] J. D. Smith, A. B. Johnson, "The impact of meeting frequency on attitudes towards meetings," *Journal of Organizational Psychology*, vol. 45, no. 2, pp. 123–137, 2018.
- [9] M. Gaborov, Ž. Stojanov, 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, 2023.
- [10] M. K. Garcia, T. H. Nguyen, "Exploring the relationship between meeting frequency and organizational communication climate: Insights from IT experts," *Communication Studies*, vol. 68, no. 4, pp. 422–437, 2017.
- [11] G. A. Bowen, "Naturalistic inquiry and the saturation concept: A research note," *Qualitative Research*, vol. 8, no. 1, pp. 137–152, 2008.
- [12] M. B. Miles, A. M. Huberman, *Qualitative Data Analysis*, 2nd ed., Sage, Thousand Oaks, CA, 1994.
- [13] K. Malterud, V. D. Siersma, A. D. Guassora, "Sample size in qualitative interview studies: Guided by information power," *Qualitative Health Research*, vol. 26, no. 13, pp. 1753–1760, 2016.
- [14] W. Daher, "Saturation in qualitative educational technology research," *Education Sciences*, vol. 13, no. 2, p. 98, 2023.
- [15] H. U. Mühlenfeld, "Differences between 'talking about' and 'admitting' sensitive behaviour in anonymous and non-anonymous web-based interviews," *Computers in Human Behavior*, vol. 21, no. 6, pp. 993–1003, 2005.
- [16] M. Q. Patton, *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*, Sage Publications, 2014.
- [17] P. Liamputtong, "Qualitative data analysis: Conceptual and practical considerations," *Health Promotion Journal of Australia*, vol. 20, no. 2, pp. 133–139, 2009.
- [18] J. Fischbach, H. Femmer, D. Mendez, D. Fucci, A. Vogelsang, "What makes agile test artifacts useful? An activity-based quality model from a practitioners' perspective," in *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pp. 1–10, 2020.
- [19] J. Bailey, "First steps in qualitative data analysis: Transcribing," *Family Practice*, vol. 25, no. 2, pp. 127–131, 2008.
- [20] W. C. Adams, "Conducting semi-structured interviews," in *Handbook of Practical Program Evaluation*, pp. 492–505, 2015.
- [21] M. M. Archibald, R. C. Ambagtsheer, M. G. Casey, M. Lawless, "Using Zoom videoconferencing for qualitative data collection: Perceptions and experiences of researchers and participants," *International Journal of Qualitative Methods*, vol. 18, p. 1609406919874596, 2019.
- [22] V. Braun, V. Clarke, *Thematic Analysis: A Practical Guide*, Sage Publications Ltd., 2022.

- [23] J. N. Lester, Y. Cho, C. R. Lochmiller, "Learning to do qualitative data analysis: A starting point," *Human Resource Development Review*, vol. 19, no. 1, pp. 94–106, 2020.
- [24] M. Denscombe, *The Good Research Guide: For Small-Scale Social Research Projects*, McGraw-Hill Education (UK), 2017.

# Pricing Dynamics in the Digital Housing Market: An Analytical Comparison of RentLingo.com and RentDigs.com

Gisele Habibulla<sup>1,2</sup>

<sup>1</sup> The Coding School

<sup>2</sup> Archbishop Molloy High School, 83-53 Manton St, Briarwood, NY

Gisele.habibulla.ii@gmail.com

## Abstract:

This paper presents a data-driven comparative analysis of two apartment listing platforms, RentLingo.com and RentDigs.com, focusing specifically on median apartment prices in New Jersey. Through systematic data collection, this study examines price variations in small-sized apartments listed across the United States, honing in on the New Jersey market. A two-sample T-test was conducted using rigorous statistical methods to evaluate the differences in median prices between the two platforms. The analysis yielded a p-value of 0.1818, suggesting no statistically significant difference in the prices of apartments listed on RentDigs.com and RentLingo.com. These findings indicate that both platforms maintain comparable pricing structures for small apartments in New Jersey. This study underscores the importance of continued exploration into other listing sources, as well as geographical and temporal factors influencing apartment pricing, to enhance our understanding of market dynamics in the digital real estate landscape.

## Keywords:

Data analysis, apartment rental platforms, pricing comparison, statistical methods, T-test, RStudio.

## Introduction

The internet has transformed many industries, including rental and real estate. Digital platforms have made it easier for people to search for rental properties, but they also introduce new challenges in terms of price comparison and market efficiency. The dataset used describes apartments listed across the United States. Data was collected on multiple apartments listed by different apartment listers, including information such as the number of bedrooms, bathrooms, size (square feet), state, time it was listed, and price. The information provided for each apartment was paired with its lister, essentially displaying a wide variety of apartments, their features, and who listed them – leading to the question: How does the apartment lister affect the price of residential apartments in similar locations?

The apartments listed were grouped by size (small 0-1300 square feet, medium 1300-4500 square feet, and large 4500+ square feet), and separated by the state they were listed in, for this experiment, we will only be looking at the apartments classified as small and located in New Jersey. This paper explores whether the platform impacts the cost of an apartment in comparable spaces. Finding an answer to this question is important for fostering market openness and supporting consumers in the online rental market.

### Research Question

How does the apartment lister affect the price of residential apartments in similar locations?

### Hypothesis

Apartments listed by RentLingo.com are more/less expensive than those listed by RentDigs.com.

Null hypothesis: the price of apartments listed by RentLingo and RentDigs are the same in similar locations.

## Literature Review

### 1.1 Implications for Market Efficiency

The digitization of the real estate rental market has made property prices more accessible and easier to compare, benefiting both consumers and the overall market. Platforms including RentLingo.com and RentDigs.com are convenient for renters to examine properties and reduce the cost/time related to searching [3]. These platforms help lessen the information asymmetry that often exists in traditional markets, allowing consumers to make more informed decisions [1]. This research reveals that although the two platforms differ slightly in pricing structure there is no significant difference in competition. By leveling the playing field, both platforms allow renters to choose between them without worrying about significant price variations.

Additionally, the competition between these platforms ensures that prices remain transparent. Renters can quickly switch between RentLingo.com and RentDigs.com to find the best deals, forcing both platforms to maintain competitive pricing [1][2]. This convergence of pricing between the two platforms indicates that neither can afford to significantly deviate from market rates without risking losing customers to the other. Glaeser and Gyourko [4] described that open housing prices create more effective results. Digital competition causes pricing on these platforms to unite, leading to fair market actions and enhancing broader economic gains from platform technology.

Moreover, the role of user experience and engagement in shaping the effectiveness of these platforms cannot be overlooked. Research by Miller and Lee [6] emphasizes that mobile applications significantly enhance user interaction, making it easier for renters to access real-time pricing data, thus further reducing information asymmetry. Similarly, Patel et al. [7] highlight that intuitive user interface design on these platforms improves customer satisfaction and increases the likelihood of successful transactions. Smith et al. [5] also notes that consumer feedback mechanisms play a vital role in influencing pricing strategies, as user reviews can drive competition and force platforms to adapt to consumer preferences. Collectively, these factors contribute to the overall efficiency of the digital rental market, reinforcing the importance of user-friendly design and real-time information access in fostering a competitive real estate market.

## Methodology

### 1.2 Data Collection

The dataset used in this research was collected from multiple apartment listing sources. Each entry in the dataset includes details such as:

- Number of bedrooms
- Number of bathrooms
- Apartment size (square feet)
- State
- Listing time
- Price

The following is a bar plot displaying the number of apartments listed by RentDigs.com and RentLingo.com by state as well as a pie chart displaying the number of apartments total listed by each source to provide some perspective on the difference in the quantity of data provided by each source. These graphs were designed to better visualize the dataset as a whole. The noticeable majority of apartments listed by RentDigs in New Jersey were taken into account, which is why this experiment focuses on specifically small apartments. In this smaller scope of apartments, they have a smaller gap in the number of apartments listed.



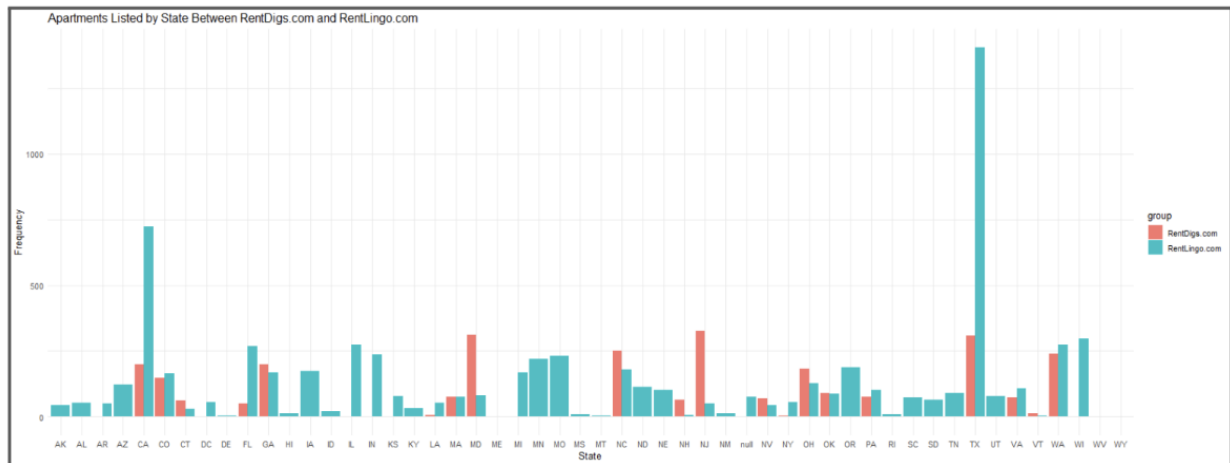


Figure 1: Bar plot demonstrating the number of apartments listed by each renting site per state.

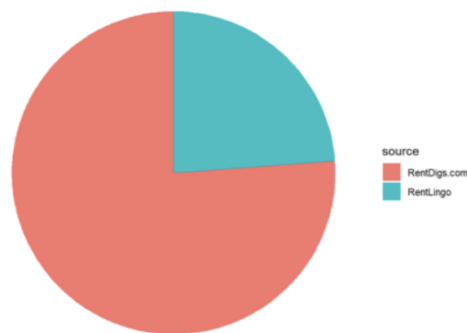


Figure 2: pie chart displaying the number of apartments total listed by each source.

### 1.3 Data Processing

The dataset was cleaned and filtered to include only listings from New Jersey with apartments classified as "small." The following R libraries were used for data manipulation and visualization:

```
library(ggplot2)
```

```
library(dplyr)
```

The data was grouped by listing source (RentLingo.com or RentDigs.com) and then organized by price for comparison.

The `ggplot()` function was used to plot the data. First, the data was organized and grouped.

```

library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)
library(janitor)
library(corrplot)

df <- read_delim("data/apartments_for_rent_classified_10K.csv", delim = ";")
rentlingodf <- filter(df, source == "RentLingo")
rentlingodfnj <- filter(rentlingodf, state == "NJ")

#Rent digs data set
rentdigsdf <- filter(df, source == "RentDigs.com")
rentdigsdfnj <- filter(rentdigsdf, state == "NJ")

rentlingodf <- filter(df, source == "RentLingo")
rentlingodfnj <- filter(rentlingodf, state == "NJ")

rentdigsdf <- filter(df, source == "RentDigs.com")
rentdigsdfnj <- filter(rentdigsdf, state == "NJ")

sources_combined <- filter(df, source %in% c("RentDigs.com", "RentLingo"))

size_grouped_df <- df %>% mutate(size_SML =
 case_when(square_feet <= 1300 ~ "small",
 square_feet <= 4500 ~ "medium",
 square_feet >= 4501 ~ "large")
)

rentlingo <- size_grouped_df %>% filter (source == "RentLingo", size_SML == "small", state == "NJ")
rentdigs <- size_grouped_df %>% filter (source == "RentDigs.com", size_SML == "small", state == "NJ")
sources_combined_2 <- size_grouped_df %>% filter (source %in% c("RentDigs.com", "RentLingo"), size_SML == "small",
state == "NJ")

merged_dataset <- rbind(
 data.frame(group = "RentDigs.com", rentdigsdf),
 data.frame(group = "RentLingo.com", rentlingodf)
)

```

## 1.4 Statistical Analysis

A two-sample T-test was then conducted to determine if there is a statistically significant difference between the median prices of apartments listed on RentLingo.com and RentDigs.com. The null hypothesis states that there is no significant price difference between the two platforms.

## Results

### 1.5 Descriptive Statistics

The table below presents summary statistics for apartment prices on the two platforms.

Platform	Mean Price (\$)	Standard Deviation (\$)	Median Price (\$)
RentLingo.com	1792.19	300.25	1800.00
RentDigs.com	1906.28	320.15	1850.00

### 1.6 T-Test Results

The T-test revealed a p-value of 0.1818, indicating that there is no statistically significant difference between the prices listed on RentLingo.com and RentDigs.com ( $p > 0.05$ ). The 95% confidence interval for the difference in means was [-282.65, 54.48], supporting the null hypothesis.

## T-Test

```
t.test(rentlingo$price, rentdigs$price)
```

```
##
Welch Two Sample t-test
##
data: rentlingo$price and rentdigs$price
t = -1.3467, df = 80.519, p-value = 0.1818
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-282.65031 54.48179
sample estimates:
mean of x mean of y
1792.191 1906.276
```

## 1.7 Data Visualizations.

### Density Plot

The following density plot visually represents the price distribution for both platforms. This plot is a visual representation of the P-value of 0.1818 (which was found in the t-test). Here, the peak of the curves represents the mean of the price data for both RentDigs.com and RentLingo.com. RentLingo may have the greater density, but the peaks are around the same price, therefore representing a relatively similar mean. The plot was generated in RStudio.

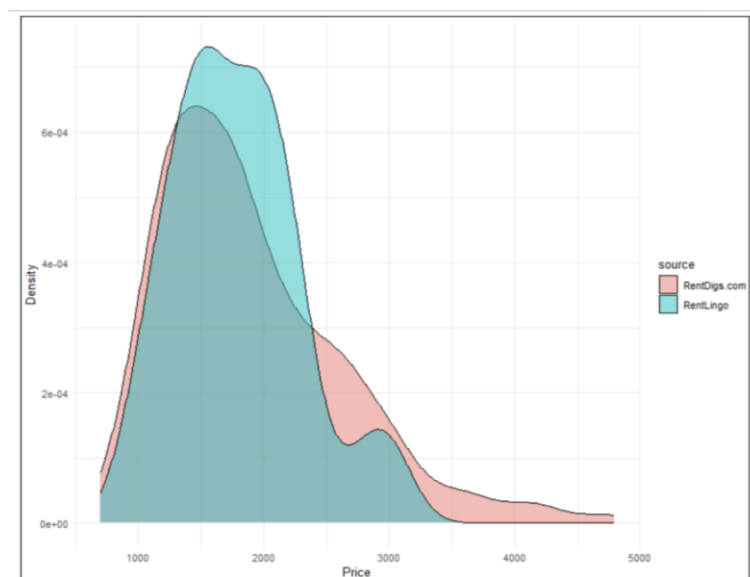


Figure 3: RentDigs and RentLingo Density vs. Price for small apartments in New Jersey.

### Box Plot

The box plot below provides an alternative representation of the price distribution. The dark line in the middle of the box represents the mean, the box itself being the interquartile range. The remaining dots show outliers. Although RentDigs, again, shows a greater distribution, both sources have relatively similar means. The plot was generated in RStudio.

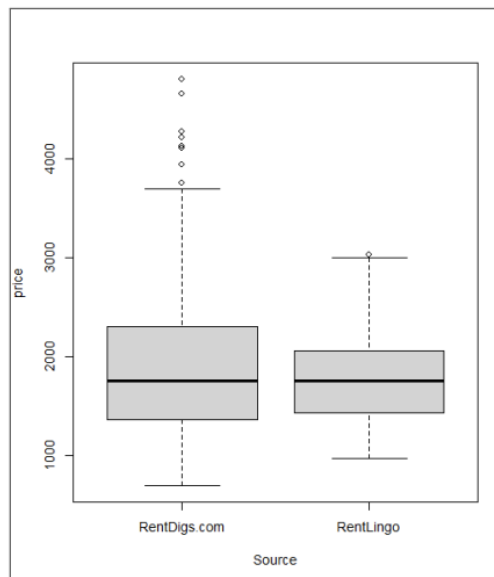


Figure 4: Price distribution between RentDigs and RentLingo for small apartments in New Jersey.

## 5. Discussion

The results fail to reject the null hypothesis, therefore proving that the prices for apartments listed in New Jersey by RentDigs.com and RentLingo.com are relatively the same regardless of the source.

$$0.1818(\text{p-value}) > 0.05$$

The results of this experiment prove that potential consumers do not need to be concerned with which listing site they use to find an apartment as the two sites show no significant differences in price. Therefore, both sites will present relatively similar prices.

### 5.1 Limitations

- RentLingo.com had a greater variety of states listed and overall more apartments listed.
- Apartments were listed at different times.
- Only two apartment listers were compared.
- RentDigs Had more apartments listed in the chosen area (by 100 square feet).

This finding is important for consumers, as it suggests that renters in New Jersey can expect to find similar prices on both platforms, regardless of which one they use. This research highlights the role of digital real estate platforms in maintaining market transparency and provides insights into their impact on consumer behavior.

### 5.2 Future Directions

Future directions for this research would be to compare the prices listed by apartments of different sizes (this experiment only compares apartments below 1400 square feet), look into the times in which the apartments were listed and how they compare. (Different times may have had different financial statuses in the chosen location when accounting for price inflation), and to build linear regression models to compare more than just 2 sources.

## 6. Conclusion

This study compared RentLingo.com and RentDigs.com for apartment listings in New Jersey. Statistical analysis in RStudio demonstrated that there is no statistically relevant difference in pricing across the two platforms, with a p-value of 0.1818. This result shows that renters can anticipate consistent pricing across both sites. The findings suggest that digital platforms enhance market transparency and effectiveness. Uniform pricing on these websites improves market transparency and competition in the rental industry, leading to fairer market conditions for consumers.

This study highlights that both RentLingo.com and RentDigs.com offer consistent prices for small apartment listings in New Jersey, giving renters confidence in receiving equal prices regardless of the platform. These findings emphasize the critical role digital real estate platforms play in enhancing market clarity and guiding consumer decisions.

Future investigations should explore pricing behaviors in different regions and apartment sizes, as well as on other listing platforms. Further analysis using linear regression models could uncover complex connections between platform competition and rental pricing. Additionally, studies could examine how factors like listing timing and larger apartment categories influence rental prices across platforms.

### Acknowledgment:

This work was made possible by the support of The Coding School's Data Science Research Program and their scholarship program. I would also like to thank Sarah Parker for her expertise, Delaney Rice for organizing the program, and Tejas Kamtam for his guidance throughout this process. Additionally, I am grateful to Dr. Di Xiong for the insightful advice and access to the data, and to my classmates for their collaborative spirit. Finally, a special thank you to Mrs. Edit Boral for her role in facilitating my involvement with The Coding School and her mentorship.

### References:

- [1] Zhan, M., & Tzeng, J. (2017). "The Role of Online Property Platforms in the Housing Market: A Case Study." *Journal of Urban Economics*, 96, 45-58.
- [2] Barlow, J., & Hennessey, A. (2021). "Consumer Behavior in the Rental Market: A Study of Online Rental Platforms." *Housing Studies*, 36(3), 422-440.
- [3] Chen, Y., & Zhang, Z. (2020). "The Impact of Online Real Estate Platforms on Housing Market Efficiency." *Real Estate Economics*, 48(4), 1025-1050.
- [4] Glaeser, E. L., & Gyourko, J. (2018). "The Economic Effects of Housing Markets." *Journal of Economic Perspectives*, 32(1), 1-30.
- [5] Smith, A., Johnson, R., & Lee, K. (2022). "Consumer Feedback Mechanisms in Digital Rental Platforms: Implications for Market Dynamics." *International Journal of Real Estate Studies*, 16(2), 115-134.
- [6] Miller, T., & Lee, S. (2023). "Mobile Applications and Consumer Engagement in the Rental Market." *Journal of Housing Research*, 32(1), 60-75.
- [7] Patel, R., Thompson, J., & Wang, Y. (2022). "The Importance of User Experience in Digital Rental Platforms." *Journal of Real Estate Literature*, 30(2), 200-215.

# From Digital Substations to Cybersecurity: Protecting the Evolving Power Transmission Network

Mugdin Agić<sup>1</sup>, Mia Lešić Aganović<sup>1</sup> and Emir Skejić<sup>2</sup>

<sup>1</sup> *Elektroprenos – Elektroprijenos BiH a.d. Banja Luka, Operativno područje Tuzla, Tuzla, B&H*

<sup>2</sup> *University of Tuzla, Faculty of Electrical Engineering, Tuzla, B&H*

*mugdin.agic@elprenos.ba; mia.lesic@fet.ba; emir.skejic@fet.ba*

## Abstract:

The integration of advanced telecommunication and digital technologies within transmission networks is essential for enhancing efficiency, reliability, and security in modern power grids. This paper proposes the approach for the digitalization of communication systems in electricity transmission networks, emphasizing the role of intelligent electronic devices (IEDs) and real-time data exchange for optimizing grid operations. By transitioning from traditional analog systems to IP-based and fiber-optic communications, transmission networks become more flexible and scalable, supporting smart grid technologies and renewable energy integration. However, this digital transformation introduces cybersecurity vulnerabilities. Key strategies for securing these critical infrastructures include implementing defense-in-depth, network segmentation, and robust identity and access management (IAM) practices. The paper also highlights the need for continuous real-time monitoring, data encryption, and employee cybersecurity awareness to mitigate threats such as malware, denial-of-service attacks, and insider threats. By adopting a multi-layered cybersecurity approach and leveraging advanced technologies like artificial intelligence and blockchain, the power transmission network can ensure secure and resilient operation.

## Keywords:

Cybersecurity, digitalization, transmission network, telecommunication system

## 1. Introduction

Computer, electronic and communication systems are an inseparable part of the transmission network as a part of electricity power system. The imperative for electric power system is to enhance information technology capabilities, along with digitalization, decarbonization, decentralization and deregulation [1, 2].

The digitalization of telecommunication systems in power transmission network refers to the integration of advanced digital technologies into the grid infrastructure. This process enhances the efficiency, reliability, and flexibility of power transmission, enabling smarter grid management and the integration of renewable energy sources. By integrating modern telecommunication technologies, transmission network is becoming more intelligent, efficient, and responsive. Contemporary automation systems for power substations are designed to gather and utilize detailed data from Intelligent Electronic Devices (IEDs). These systems aim to facilitate dependable and effective remote management of substation components and an equipment from operational control centers and data hubs, ensuring consistent and reliable communication between them [3-5]. Electricity transmission network is often set up by multiple operational areas due to its wide geographical distribution. To improve system management, it is necessary to adopt digitalization and virtualization technology to build flexible resource management. Communication technologies for electricity transmission networks, along with the key differences between Supervisory Control and Data Acquisition (SCADA) systems and traditional IT systems, are described in detail in [6].

Digital substation is one with advanced automation, where nearly all aspects of information transfer between substation elements, interaction with external systems, and management of the substation's operation are conducted through digital means [7]. Digital substation is a term that refers to electrical

substations where operations are managed through a network of distributed IEDs connected by a communication network [8].

The impact of the security breaches in power substations must not be underrated. Emerging cybersecurity issues in power substations must be effectively managed [9-12]. The review of cybersecurity studies in electric power systems is presented in [13], while potential threats are described in [14, 15].

The increasing reliance on digital communication systems also introduce vulnerabilities, making cybersecurity a critical concern. As power transmission network is vital to national infrastructure, ensuring their security against cyber threats is paramount. Ensuring cybersecurity in transmission network substations is a complex problem. Safety standards must be met and implemented in such way that the fulfillment of safety requirements does not affect the operation of the power system [16].

This paper explores the key cybersecurity strategies necessary to protect digitalized telecommunication systems. The paper addresses a critical gap by discussing both benefits and the cybersecurity vulnerabilities that arise with digitalization. The most important contributions of proposed approach in this paper are:

- A comprehensive overview of the digitalization process in transmission networks and its impact on operational efficiency and flexibility.
- Detailed insights into the cybersecurity risks introduced by digitalization, such as malware, APTs and DoS attacks.
- Recommendation for using advanced technologies, including machine learning, artificial intelligence (AI) and blockchain, to enhance threat detection and automate cybersecurity defenses.

To present the research, this paper is organized into three sections that explore the process of the digitalization of telecommunication systems within the power transmission network and the associated cybersecurity challenges. The introductory part outlines the importance of communication, infrastructure in power grids and introduces the need for digitalization, decarbonization, decentralization and cybersecurity measures. The second section deals with the technological transformation in power grids, highlighting the shift from traditional analog systems to modern IP based communication networks. Cybersecurity strategies follows as third section, focusing on the vulnerabilities introduced by digital systems and proposing key defenses such as defense-in-depth, IAM, network segmentation, encryption, and real-time monitoring.

## 2. Digitalization of communication system within transmission network

Maintaining a constant balance between electricity demand and supply is highly complex due to several factors. These include the geographic separation of electricity production and consumption across different regions, the partial predictability of user consumption profiles, the significant impact of weather conditions on production, particularly from renewable energy sources, and the need to manage the electricity grid to prevent overloads and handle outages. This process, known as dispatching, is managed by national Transmission System Operators (TSOs). TSOs must continuously oversee electricity flows to ensure that the balance between available and consumed electricity is maintained at every point in the network. A schematic representation of the transmission network as a part of power system is illustrated in Figure 1.

Traditionally, power grids relied on analog communication systems, such as Power Line Carrier (PLC) and microwave communication, which are limited in bandwidth and flexibility. The supervision and control of power systems are carried out through dispatch centers. For communication, the substations are connected by an optical ground wire (OPGW) by means of synchronous and plesiochronous digital hierarchy (SDH/PDH) devices [17]. Network management systems (NMS) have been implemented to enable the control and access to each individual network element and a management of the entire network from a central location at any time. With digitalization, fiber optic networks, IP-based communication, and wireless technologies such as LTE and 5G are increasingly used. These systems provide higher bandwidth, lower latency, and more reliable communication channels. Transmission network now needs a communication that is more flexible and can accommodate rapid additions and changes to nodes and connections. Internet Protocol/Multiprotocol

Label Switching (IP/MPLS) is able to ensure quality of service, scalability, availability, reliability, and security. It allows a transmission of different types of the communication, using virtual paths [18].

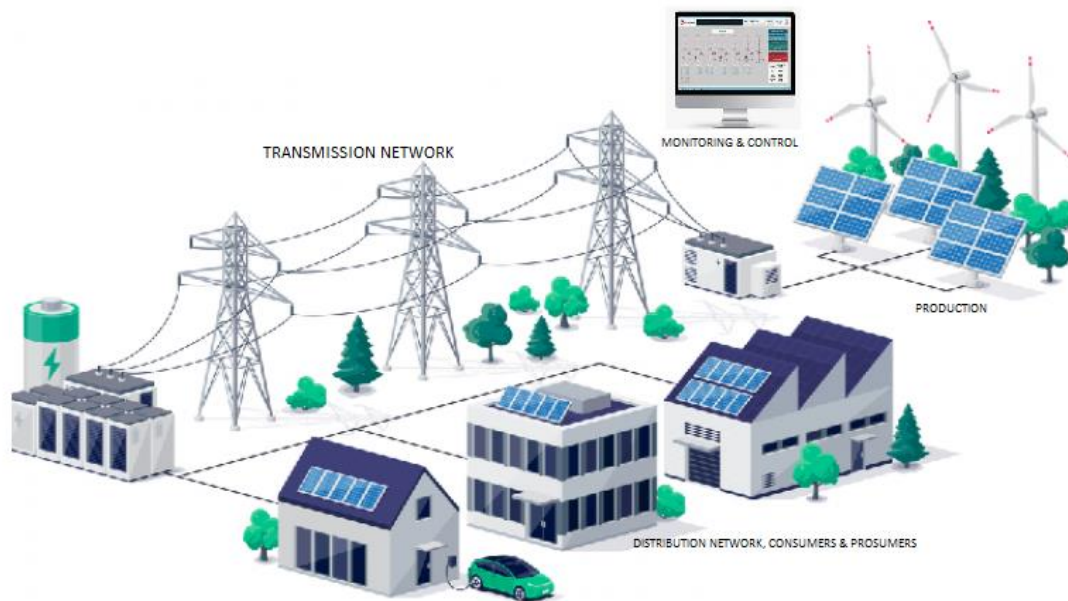


Figure 1. Schematic representation of power system

A substation in transmission network consists of closely connected subparts with some common functionality. In the electricity transmission system substations, devices receive IP protocols for access and monitoring. At the lower level of operation, there is not enough range of IP addresses for multiple devices and software that can be remotely monitored and managed (except for the SCADA system). Thus, maintenance and exploitation services include procedures for visiting substations from headquarters of the operational area even for the smallest adjustments and management of certain devices, creating unnecessary business costs. The newer generation of devices and software in substations enables their connection and exploitation through the LAN network. These devices include electricity meters, protection and control devices, inverters/rectifiers/UPSs, alarm devices, video surveillance devices, records of working hours, computers with management and monitoring software, telephony devices, and others. The communication infrastructure commonly used in electrical networks is illustrated in Figure 2 [19]. A key feature of a grid operator's network is the distinct separation between the office network and the process control network [19]. This separation is crucial for maintaining the security and reliability of critical infrastructure operations. The office network operates similarly to standard corporate IT systems, primarily handling business operations, administrative tasks, and non-critical applications. In contrast, the process control network links the dispatch center to substations and field devices, prioritizing real-time control, monitoring, and automation of the electrical grid.

The control center generally includes a human-machine interface (HMI), a database server for managing grid data, and a simulation server used for forecasting the effects of grid adjustments. Moreover, it is connected to various substations, each of which generally includes a gateway, an HMI, and several PLCs, allowing for communication with other SCADA systems from Transmission System Operators (TSOs) for coordinated control. To manage data flow between the office and process control networks, a dedicated data exchange server is required, ensuring all files are scanned for malware before transmission.

Digital communication enables real-time monitoring and control of HV substations and transmission lines. This helps in early fault detection, reducing downtime and improving overall grid reliability. Digitalization facilitates the adoption of smart grid technologies like Advanced Metering Infrastructure (AMI) and Distribution Management Systems (DMS), promoting more efficient energy distribution and improved demand response. It also allows more sophisticated protection schemes, enabling faster and more coordinated responses to grid disturbances. High-speed data transmission enables the use of



big data analytics and artificial intelligence for predictive maintenance, load forecasting and optimizing grid operations.

Ensuring that different systems and devices can communicate effectively requires an adherence to standards and protocols, which can be complex and costly. Transitioning from legacy systems to digital communication requires a significant investment in new infrastructure and technology, along with training for personnel [20].

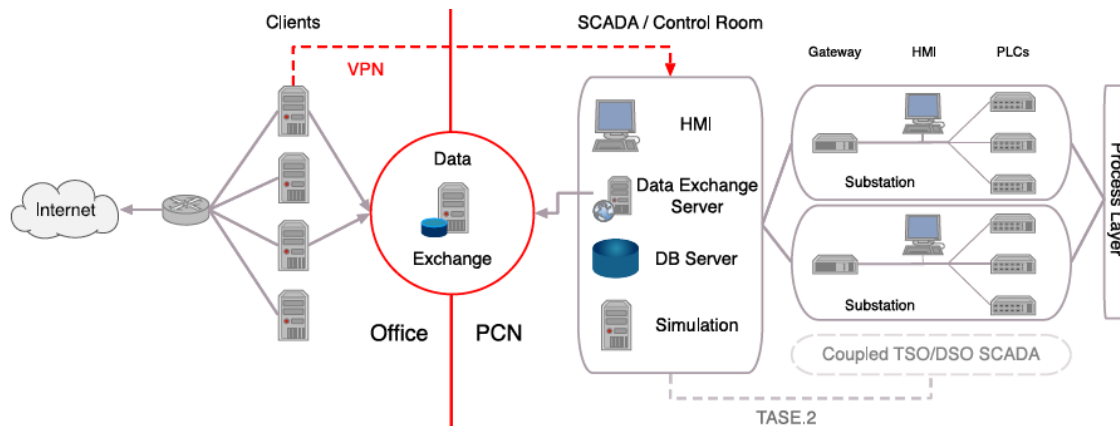


Figure 2. Simplified view of TSO network

### 3. Cybersecurity strategies in the digitalization of telecommunication systems in transmission network

As digital technology continues to advance, the demand for enhanced network capabilities has increased, linking previously isolated components into expansive communication networks. This expansion leads to a wide array of new data flows [19]. Additionally, the growing interconnection of power grids, which are vital infrastructure, introduces substantial security concerns. Cybersecurity strategies in the digitalization of telecommunication systems within transmission networks are crucial to safeguarding these critical infrastructures from a growing array of cyber threats [21]. As these networks become more dependent on digital technologies, the attack area expands, making them more vulnerable to cyber-attacks. Before diving into specific strategies, it is essential to understand the potential threats to high-voltage power transmission networks [7, 22]:

- Advanced persistent threats (APTs): cyber adversaries may target power grids with sophisticated, long-term attacks aimed at compromising critical systems.
- Malware and ransomware: Malicious software can infiltrate network devices, disrupt operations, or hold critical systems hostage.
- Denial-of-service (DoS) attacks: These attacks overwhelm communication networks, causing disruptions in data transmission and potentially leading to power outages.
- Insider threats: Employees or contractors with access to the network might unintentionally or maliciously compromise security.

Given these threats, a robust multi-layered cybersecurity strategy is essential. Digital systems are vulnerable to cyber-attacks that could disrupt power transmission and lead to widespread outages. Implementing strong cybersecurity measures is crucial. The main security challenge in electrical networks arises from the digitalization of communication and the reliance on common Ethernet technologies. Although this approach simplifies design, it also heightens vulnerability to standard attacks [23]. Additional challenges include inadequate credential validation, insufficient encryption and data verification, poor credential management, and issues with software configuration, maintenance, and access control. According to [8], a thorough security analysis for digital substations should encompass the security pyramid, threat analysis, threat modeling, risk assessment, and impact analysis through simulations.

### 3.1. Implementing a defense-in-depth approach

A defense-in-depth strategy employs multiple layers of security measures to safeguard the network from different angles, as shown in Figure 3. This approach ensures that even if one layer is breached, others will still provide protection. Key components include:

- **Network segmentation:** By dividing the network into segments, critical systems can be isolated from less secure parts, limiting the spread of potential breach. It is crucial to maintain a clear separation between operational technology (OT) and information technology (IT) systems, which handle data and communication. This minimizes the risk of an attack spreading across entire network.
- **Firewall and intrusion detection/prevention systems (IDS/IPS):** IDS/IPS monitor and control a network traffic, blocking unauthorized access and detecting suspicious activity. Deploying firewalls at critical points within the network is necessary to control and monitor incoming and outgoing traffic and helps to block unauthorized access. An Intrusion Detection System (IDS) monitors network traffic for unusual activities and notifies administrators when potential threats are identified. IPS not only detect but also actively block or mitigate identified threats in real-time.
- **Data encryption:** Encrypting data both at rest and during transit ensures that, even if intercepted, it cannot be easily accessed or used by unauthorized parties. The data remains inaccessible without the correct decryption keys. End-to-end encryption strengthens security by ensuring that data remains encrypted throughout its entire transmission, from the source to the destination.

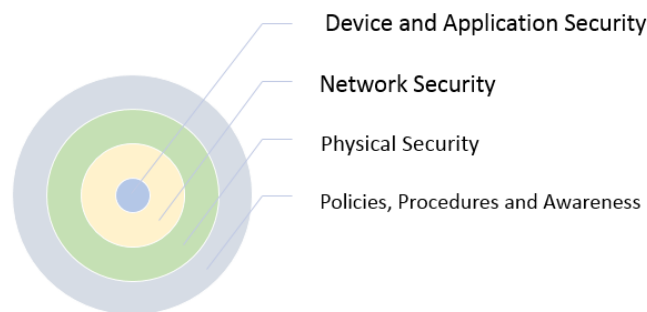


Figure 3. Defense-in-depth principle

### 3.2. Strengthening an identity and access management

Controlling who has access to the network and what they can do is fundamental aspect of cybersecurity. Key IAM strategies include [24]:

- **Multi-factor authentication (MFA):** Implementing multiple forms of verification for access minimizes the risk of unauthorized entry, even if passwords are compromised. This significantly enhances security by requiring additional proof of identity.
- **Role-based access control (RBAC):** Restricting access based on a user's role ensures that employees can only access the systems and data needed for their job functions. This approach helps to limit potential damage in the event of a security breach.
- **Regular audits and reviews:** Periodic audits of access rights help to identify and revoke unnecessary privileges, reducing potential insider threats or compromised accounts being used maliciously.

### 3.3. Enhancing a cyber threat and detection response

Continuous monitoring and a well-established incident response plan are essential for promptly detecting and addressing to cyber threats. These involve:

- **Real-time monitoring:** Implementing systems that provide real-time visibility into network activity helps in early detection of anomalies or breaches. Monitoring network traffic patterns can reveal signs of an attack, such as unusual data flows or communication with known malicious IP addresses.
- **Security information and event management (SIEM):** SIEM systems collect and analyze data from multiple sources to provide insights into potential security incidents, enabling a swift and effective response.
- **Incident response plan (IRP):** An effective IRP outlines procedures for containing, mitigation and recovering from a cyber incident. A well-practiced IRP helps minimize the impact of an attack and ensures a swift return to normal operations.

### 3.4. Building a culture of cybersecurity awareness

Human error often significantly contributes to cybersecurity breaches. Historically, the most severe cyber-attacks on electrical networks have exploited lapses in human actions, often through spear-phishing, especially via emails or manipulated downloads [25]. Establishing a culture of cybersecurity awareness within the organization is crucial:

- **Training programs:** Regular cybersecurity training helps employees to identify phishing attempts, understand the importance of secure practice, and know how to report suspicious activities.
- **Phishing simulations:** Running simulated phishing attacks can evaluate employees' preparedness and reinforce lessons learned from training.

### 3.5. Investing in cybersecurity technologies

As cyber threats evolve, investing in advanced cybersecurity technologies is of essential importance:

- **Artificial Intelligence [26] and Machine Learning (ML) [27]:** AI and ML algorithms are capable of analyzing large amounts of data to identify patterns and anomalies that may indicate a cyber-attack. These technologies offer faster and more precise threat detection compared to traditional methods. Additionally, they can automate responses to specific types of threats, thereby shortening the time between detection and mitigation.

AI and ML offer promising tools for a threat detection and an automation but face significant challenges in transmission networks such as requiring large volumes of real-time, high-quality data, continuous updates and specialized expertise and high computational power. Also, false positive and negative system signals can disrupt grid operations or fail to detect actual threats, making it crucial to balance precision and sensitivity.

- **Blockchain technology [28]:** The blockchain can be used to create immutable records of transactions or communications within the network, ensuring that data cannot be tampered without detection. A blockchain can also enhance IAM by providing a decentralized and secure way of managing user identities and access rights. The blockchain also currently presents several challenges in terms of scalability, energy consumption, inoperability, and privacy.

### 3.6. Regular testing and updating of security measures

Cybersecurity is a dynamic field, and regular testing and updates are necessary to maintain robust defenses:

- **Penetration testing:** Regular penetration testing, where ethical hackers attempt to breach the network, helps identify vulnerabilities before malicious actors can exploit them. Vulnerability assessments provide insights into the weaknesses within the network and guide an implementation of necessary security improvements.
- **Patch management:** Keeping all software and hardware systems updated with the latest security patches is essential for addressing known vulnerabilities that could be exploited by attackers. Utilizing automated patch management systems helps to ensure that patches are applied swiftly across the network, minimizing the risk of exposure to potential threats.

## 4. Conclusions

Digital systems enable faster fault detection and isolation, reducing a duration and an impact of outages. Better data collection and analysis lead to more efficient grid operation, reducing losses and optimizing power flows. Digital communication systems can easily scale to accommodate future grid expansions and the increasing complexity of modern power systems.

In short, the digitalization of telecommunication systems in high-voltage transmission networks is crucial step toward building smarter, more efficient, and resilient power grids. It involves the adoption of advances communication technologies, an adherence to industry standards and a focus on cybersecurity to address the challenges of modern energy demands. The digitalization of telecommunication systems in high-voltage power transmission networks offers tremendous benefits, but also introduces significant cybersecurity challenges. By implementing a comprehensive, multi-layered cybersecurity strategies that include a defense-in-depth, an adherence to industry standards, robust IAM practices, continuous monitoring, and a culture of awareness, the organizations can significantly mitigate these risks.

This paper highlights how advanced technologies, such as artificial intelligence, machine learning, and blockchain, offer promising solutions for automated threat detection, response, and securing grid operations. However, their successful implementation requires further research to overcome technical challenges and scalability issues in transmission networks.

The strategies to protect these critical infrastructures, which ensure the secure and reliable operation of power transmission networks now and in future, must follow recent advances in the technology. Looking forward, further research is necessary to explore the potential of advanced artificial intelligence and machine learning algorithms for predictive threat detection and automated cybersecurity responses. Additionally, blockchain technology could be leveraged for secure transaction management and an identity protection within decentralized power networks. As cyber-physical systems become increasingly interconnected with the grid, securing these systems against sophisticated cyber-attacks will be crucial. Moreover, with the advent of quantum computing, it is essential to develop quantum-resistant encryption protocols to future-proof communication in transmission networks. Finally, digital twin technology holds promise for simulating grid operations and cyber-attacks, providing real-time tools for risk management and enhancing grid resilience.

By addressing these areas, the security and operational efficiency of digitalized transmission networks can be strengthened, ensuring a reliable and sustainable energy future.

### References:

- [1] Y. Zeng, Revolutionizing Power Systems Through Electrical Automation: Efficiency, Reliability and Sustainability, Academic Journal of Science and Technology, Vol. 10, No. 1, 2024.
- [2] F. Heymann, T. Milojevic, A. Covatariu, P. Verma, Digitalization in decarbonizing electricity systems - Phenomena, regional aspects, stakeholders, use cases, challenges and policy options, Elsevier, Energy, 2023.
- [3] M. Z. Gunduz, R. Das, Cyber-security on Smart Grid: Threats and Potential Solutions. Comput. Netw. 2020.
- [4] W. Pavon, E. Inga, S. Simani, M. Nonato, A Review on Optimal Control for the Smart Grid Electrical Substation Enhancing Transition Stability. Energies 2021.
- [5] F. E. Abrahamsen, Y. Ai, M. Cheffena, Communication Technologies for Smart Grid: A Comprehensive Survey, Sensors, 2021.
- [6] S. Chountasis: Communication Challenges for Electric Power Transmission Systems, International Journal of Engineering and Technologies, ISSN: 2297-623X, vol 21, pp 1-12, 2021.
- [7] I. Kolosok, E. Korkina. Problems of Cyber Security of Digital Substations. In VIth International Workshop Critical Infrastructures: Contingency Management, Intelligent, Agent-Based, Cloud Computing And Cyber Security, pp. 75-78, Atlantis Press, 2019.
- [8] S. Yildirim Yayilgan, F. Holik, M. Abomhara, D. Abraham, A. Gebremedhin. An approach for analyzing cyber security threats and attacks: a case study of digital substations in Norway, Electronics, 11(23), 2022.

- [9] IEEE Standard Cybersecurity Requirements for Substation Automation, Protection, and Control Systems, IEEE Power and Energy Society, New York, USA, 2015.
- [10] J. Horalek, V. Sobeslav: Security Baseline for Substation Automation Systems, Sensors, 2023.
- [11] M.A. Aftab, S. M. S. Hussain, I. Ali, T.S. Ustun, IEC 61850 based substation automation system: a survey, Int. J. Electr. Power Energy Syst. 120, 2020.
- [12] CIGRE WG B5, Experience Concerning Availability and Reliability Digital Substation Automation Systems (DSAS), 42, CIGRE, 2017.
- [13] C. C. Sun, A. Hahn, C. C. Liu, Cyber security of a power grid: state-of-the-art, Int. J. Electr. Power Energy Syst. 99, 2018.
- [14] S. E. Quincozes, C. Albuquerque, D. Passos, D. Mosse, A survey on intrusion detection and prevention systems in digital substations, Computer Networks, 184, 2021.
- [15] S. M. S. Hussain, T.S. Ustun, A. Kalam, A review of IEC 62351 security mechanisms for IEC 61850 message exchanges, IEEE Transactions on Industrial Informatics, 16, 2020.
- [16] J. Horalek, V. Sobeslav, Security Baseline for Substation Automation Systems. Sensors, 23(16), 2023.
- [17] I. Tadić, Availability of Links in the SDH Network EES BiH, INFOTEH-JAHORINA Vol. 12, 2013.
- [18] R. Walton, The future grid wavelength: Migrating SDH to Internet Protocol, Power Engineering, 2021.
- [19] T. Krause, R. Ernst, B. Klaer, I. Hacker, M. Henze. Cybersecurity in power grids: Challenges and opportunities. Sensors, 21(18), 2021.
- [20] E. Skejić, O. Džindo, D. Demirović, Virtualization of hardware resources as a method of power savings in data center. In The 33rd International Convention MIPRO (pp. 636-640). IEEE, 2010.
- [21] R. Setola, F. Morelli, Cyber Security Strategies for the Protection of Electrical Substations. In ITASEC, pp. 195-206, 2022.
- [22] J. Hong, C. C. Liu, M. Govindarasu, Detection of cyber intrusions using network-based multicast messages for substation automation. In ISGT 2014, IEEE, 2014.
- [23] F. Holik, S. Yayilgan, G. B. Olsborg, Emulation of Digital Substations Communication for Cyber Security Awareness. Electronics, 13(12), 2024.
- [24] N. Akyeşilmen, A. Alhosban, Non-Technical Cyber-Attacks and International Cybersecurity: The Case of Social Engineering. Gaziantep University Journal of Social Sciences, 23(1), 342-360, 2024.
- [25] S. Aboukadri, A. Ouaddah, A. Mezrioui, Machine learning in identity and access management systems: Survey and deep dive. Computers & Security, 2024.
- [26] N. Yalçın, S. Çakır, S. Üaldı, Attack Detection Using Artificial Intelligence Methods for SCADA Security. IEEE Internet of Things Journal, 2024.
- [27] I. Šenk, S. Tegeltija, L. Tarjan, Machine Learning in Modern SCADA Systems: Opportunities and Challenges, 23rd International Symposium INFOTEH-JAHORINA, 2024.
- [28] A. Augello, P. Gallo, E. R. Sanseverino, G. Sciumè, M. Tornatore, A coexistence analysis of blockchain, SCADA systems, and Openadr for energy services provision. IEEE Access, 2022.

# Machine learning and theory of information in natural language processing

Dejan Djukic<sup>1</sup>, Zlatko Radovanovic<sup>1</sup>

<sup>1</sup> Alfa BK University, Palmir Togliati 3, Belgrade, Serbia

*dejan.djukic@alfa.edu.rs; zlajo@pobox.sk*

## Abstract:

Certain methods of artificial intelligence have been used for generating natural-like human language in applications. However, there is a growing realisation that, in addition, the methods of linguistics need to be applied. The necessity of establishing a two-way communication between humans and computers initiated the creation of a new category of problems: automated generation of human language: NLG (Natural language generation). In this work, NLG has been achieved through the use of Markov chains. The parameters of a Markov chain have been estimated from a natural language text, and then text fragments have been generated using the pre-trained Markov chain.

## Keywords:

natural language processing, automatic text generation, machine learning, Markov chains

## 1. Introduction

One of the founders of the field of theory of information, Claude Elwood Shannon, has applied the methods of his newly established techniques to the automated generation of texts. In [2], Shannon describes production of random sequences of various orders. Generating strings of random words of zero-order produces incomprehensible sentences. However, when the strings of words are being generated according to their frequency in longer sequences of words appearing in the natural language corpus, this automatically generated text takes more and more natural-like form. These sub-sequences of  $n$  words ( $n$ -grams) can be viewed as transition paths from one word to the next.

On another hand, Russian mathematician Markov had also been interested in problems such as: What is the probability of a particular vowel appearing after a given vowel in Pushkins Eugene Oniegin. Some examples of the statistical analysis of sampling and chaining elements of text of Eugene Oniegin have been published in [3].

Natural language generation (NLG), as has been defined for the domain of Artificial Intelligence (AI) as a part of Natural language processing (NLP), which has the aim of generating sensible phrases and sentences appearing as natural language. NLG is essentially an automated process for generating discourse, however, whilst NLG can "write" it cannot "read". This task is being solved by the methods of Natural language understanding (NLU). Both NLG and NLU are sub-domains of NLP, which embrace both interpretation and generation of text, whether spoken or written:

- NLU is concerned with the semantical content of the natural language by taking into account the grammatical rules, the context, and the intent.
- NLP, taken in the narrowest sense, transforms a natural language discourse, a text, into predetermined language data structures.
- NLG generates natural-like discourse, text or speech, from the determined language data structures.

Thus NLG uses the recorded data, and through processes of filtering and transforation, it generates all kinds of written texts, such as reports, minutes from meetings, notes from lectures and seminars, and many more. At the same time, the theoretical advances in NLG make it a valuable tool not only in IT and computer domain, but also in psycho-linguistics and cognitive sciences [5].

In the body of this article, a software for estimating parameters of Markov chains, which has been written purposefully for this project, has been applied to a natural text. Then, word sequences have

been generated by the pre-trained Markov chain. The software and the examples of its application are presented below.

## 2. NLG models and tools

NLG, as a part of NLP, and more broadly, a part of AI, makes use of a number of algorithms and methods in order to resolve problems of generating natural-like text [4][5][9].

- **Markov chains** Markov chains have been one of the first methods used for generating natural language. The model used here predicts the next word in a sequence from the previous word and taking into account the probabilities of two or more words being in close relation. An example of the use of Markov chains has been in keyboard applications in smart phones, assisting writing by producing word suggestions.
- **Recurrent neuronal networks (RNN)** Neuronal networks are computational models attempting to imitate the working of the human brain. In RNN, each member of a data sequence is being processed during the forward flow of the network, the result of which is fed back to the network input to be used in addition to the next member of the sequence. In each iteration, the model stores previously occurring words and computes the probability of the occurrence of the next word. The existence of the "memory" makes this model almost ideal for generating text, because it can keep the context. However, as the word sequence becomes longer, RNN becomes incapable of memorising distant words. Due to this drawback, RNN cannot produce long coherent sentences.
- **Long short term memory (LSTM)** In order to resolve the issue of distant word interdependence, a variant of RNN, Long short term memory, LSTM, has been introduced. Similar to RNN, LSTM comprises four layers. These layers are: the memorising core unit, the input gate, the output gate, and the forgetting gate. LSTM can memorise or forget words in any time interval by adapting the data updating the core. At the end of the sentence, the forgetting gate becomes aware of the arrival of the switching of the context, and can cause for the current state of the core to be ignored. The network keeps track only of relevant data, and, at the same time, reduces the problem of disappearing gradients, which enables long term memory. Nevertheless, the memorising capacity of LSTM is limited to about a few hundreds of words, because of the inherently complex sequential paths from a previous unit to the next one. This complexity results in high computational demands, which hampers the learning process and parallelisation.
- **Transformer** This relatively new model has been presented in 2017 in Google's document Awareness is all you need, in which a new method of "self help mechanism" has been proposed. The Transformer comprises a set of encoders for inputs of arbitrary lengths, and another set of encoders for generating output sentences. Contrary to LSTM, the Transformer performs only a small constant number of steps, whilst the self-help mechanism directly simulates the relations between all the words in a sentence. In opposition to the previous models, the Transformer implements a representation of all the words in a context, without a necessity to compress all the information to a unique representation of a fixed length, which enables the system to handle long sentences without exorbitant increase in computing capacity. One of the most known examples of the Transformer is the GPT-2 language model by OpenAI.

## 3. Markov chains and text generation

Markov chains are present in many applications, from predictive processes to biology, weather forecasting, etc. and they have proved to be versatile tools. Some of the uses of Markov chains are:

- Text generation (the subject of this work).
- Financial modelling and forecasting, including trading algorithms
- Logistics, and modelling of future deliveries and trips
- Web searching - page ranking may be viewed as modelling random web surfing with a Markov chain

It may be concluded that this algorithm is universally useful.

### 3.1. Essential description of Markov chains

Markov chains are relatively simple way of statistical modelling of random processes [3]. They are applied in various domains, from text generation to financial modelling. They are conceptually quite intuitive, and accessible, and may be implemented without the use of advanced statistical or mathematical methods.

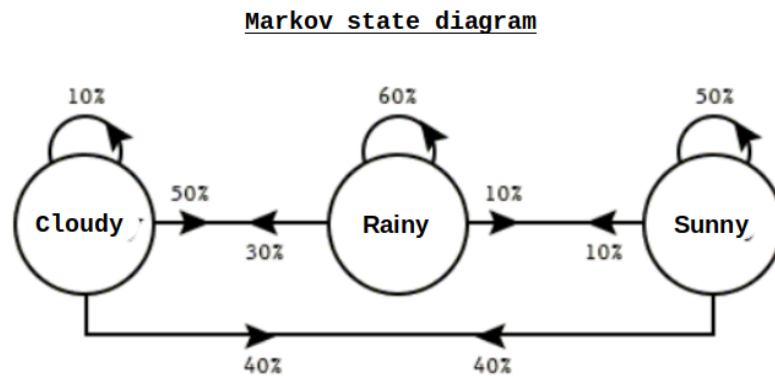


Figure 1: Markov State Diagram

#### Markov Table of Probabilities

state	next state	probability	%
Cloudy	Cloudy	0.1	10 %
Cloudy	Rainy	0.5	50 %
Cloudy	Sunny	0.4	40 %
Rainy	Cloudy	0.3	30 %
Rainy	Rainy	0.6	60 %
Rainy	Sunny	0.1	10 %
Sunny	Cloudy	0.4	40 %
Sunny	Rainy	0.1	10 %
Sunny	Sunny	0.5	50 %

Figure 2: Probabilities of weather transitions

#### Transition matrix

	Cloudy	Rainy	Sunny
Cloudy	0.1	0.5	0.4
Rainy	0.3	0.6	0.1
Sunny	0.4	0.1	0.5

Figure 3: Matrix of transition probabilities

Let us suppose that there are three possible states of weather: sunny, clouded, rainy ( **Figure 1** ). The current state of the weather is always directly observable, and its value will always be one of the



three enumerated above. The weather transitions thus form a finite state system, however, this system is not deterministic. It is stochastic. The state transitions are governed by probabilities. In order to predict tomorrow's weather, the daily transitions are observed over some long time period, and their frequencies are noted ( **Figure 2** ). Hence, the probabilities of the transitions may be computed. These probabilities may be represented by a matrix of probabilities of state transitions ( **Figure 3** ).

$$\begin{array}{c}
 \begin{array}{ccc} & \text{c}l\text{o}u\text{d}y & \text{R}a\text{i}n\text{y} & \text{S}u\text{n}n\text{y} \\ \hline \text{c}l\text{o}u\text{d}y & 1 & 0 & 0 \\ \text{R}a\text{i}n\text{y} & 0 & 1 & 0 \\ \text{S}u\text{n}n\text{y} & 0 & 0 & 1 \end{array} & \times & \begin{array}{ccc} & \text{c}l\text{o}u\text{d}y & \text{R}a\text{i}n\text{y} & \text{S}u\text{n}n\text{y} \\ \hline \text{c}l\text{o}u\text{d}y & 0.1 & 0.5 & 0.4 \\ \text{R}a\text{i}n\text{y} & 0.3 & 0.6 & 0.1 \\ \text{S}u\text{n}n\text{y} & 0.4 & 0.1 & 0.5 \end{array} \\
 \\
 = & & \begin{array}{ccc} & \text{c}l\text{o}u\text{d}y & \text{R}a\text{i}n\text{y} & \text{S}u\text{n}n\text{y} \\ \hline & 1 \times 0.1 + 0 \times 0.3 + 0 \times 0.4 & 1 \times 0.5 + 0 \times 0.6 + 0 \times 0.1 & 1 \times 0.4 + 0 \times 0.1 + 0 \times 0.5 \\ \hline & 0.1 & 0.5 & 0.4 \end{array}
 \end{array}$$

Figure 4: Example of a weather forecast

With the help of this matrix, it is possible to compute a probability of the state of the weather for tomorrow by multiplying the transition matrix with the vector state of today's weather as shown in **Figure 4** . Indeed, it is possible to compute the probability of the state of the weather for some number of days in the future, by performing this multiplication sequentially certain number of times [3][6][8][11][13].

### 3.2. Modelling text generation by Markov chains

Natural language model using Markov chains has been implemented in software using Python programming language. For this, Natural Language Toolkit (NLTK) software package has been used. The model has been implemented as two procedures. The first procedure ( **Table 1** ) generates the language model from the sample text. This programme computes the probabilities of occurrences of 2-grams, 3-grams, and 4-grams [1][4][5][7][9][10][12]. The computed probabilities have been saved in corresponding **.pkl** files.

**Table 1:**

Python code for estimating parameters of the Markov chain from text

---

```

import pickle
from collections import Counter
from nltk import ngrams
from nltk.tokenize import word_tokenize

with open("data/clean/text.txt", "r") as infile:
 text = infile.read().split("\n")

tokens = []
bigrams = []
trigrams = []
fourgrams = []
line_length = []

for line in text:
 line = line.lower()
 tkns = word_tokenize(line)
 line_length.append(len(tkns))

```

---

---

```

tokens.extend(tkns)
bgrms = list(ngrams(tkns, 2))
bigrams.extend(bgrms)
trgms = list(ngrams(tkns, 3))
trigrams.extend(trgms)
frgrms = list(ngrams(tkns, 4))
fourgrams.extend(frgrms)
count frequencies
token_freq = Counter(tokens)
bigram_freq = Counter(bigrams)
trigram_freq = Counter(trigrams)
four_freq = Counter(fourgrams)

token_total = sum(token_freq.values())
bigram_total = sum(bigram_freq.values())
trigram_total = sum(trigram_freq.values())
four_total = sum(four_freq.values())

relative frequencies
token_rel_freq = {k: (v / token_total) + 1 for (k, v) in token_freq.items()}
bigram_rel_freq = {k: (v / bigram_total) + 1 for (k, v) in bigram_freq.items()}
trigram_rel_freq = {k: (v / trigram_total) + 1 for (k, v) in trigram_freq.items()}
four_rel_freq = {k: (v / four_total) + 1 for (k, v) in four_freq.items()}

with open("models/tokens_text.pkl", "wb") as outfile:
 pickle.dump(token_rel_freq, outfile)

with open("models/bigrams_text.pkl", "wb") as outfile:
 pickle.dump(bigram_rel_freq, outfile)

with open("models/trigrams_text.pkl", "wb") as outfile:
 pickle.dump(trigram_rel_freq, outfile)

with open("models/fourgrams_text.pkl", "wb") as outfile:
 pickle.dump(four_rel_freq, outfile)

print(f"The average line length is {sum(line_length)/len(line_length)}")

```

---

Random text is generated by the procedure in **Table 2**. This procedure takes three arguments: the corpus, the initial word, and the length of sequences from which the probabilities have been computed.

**Table 2:**

Python code for generating word sequences using Markov chain

---

```

import pickle
import random

from mosestokenizer import MosesDetokenizer

text
with open("models/tokens_text.pkl", "rb") as infile:
 tokens_text = pickle.load(infile)

with open("models/bigrams_text.pkl", "rb") as infile:
 bigrams_text = pickle.load(infile)

with open("models/trigrams_text.pkl", "rb") as infile:
 trigrams_text = pickle.load(infile)

with open("models/fourgrams_text.pkl", "rb") as infile:
 fourgrams_text = pickle.load(infile)

the average line length is about 7
def simple_generator(corpus, seed, n):
 assert n in [2, 3, 4]
 tokens = tokens_text
 bigrams = bigrams_text
 trigrams = trigrams_text
 fourgrams = fourgrams_text

```

---

---

```

if seed not in tokens:
 seed = random.choice(list(tokens.keys()))
if n == 2:
 bgms = {k: v for (k, v) in bigrams.items() if k[0] == seed}
 wds = [e[1] for e in bgms.keys()]
 if wds:
 weights = [float(e) for e in bgms.values()]
 return random.choices(population=wds, weights=weights)[0]
 else:
 return random.choices(population=list(tokens.keys()),
 weights=list(tokens.values()))[0]
elif n == 3:
 tgms = {k: v for (k, v) in trigrams.items() if k[:2] == seed}
 wds = [e[2] for e in tgms.keys()]
 if wds:
 weights = [float(e) for e in tgms.values()]
 return random.choices(population=wds, weights=weights)[0]
 else:
 w = random.choice(list(tokens.keys()))
 return w
else:
 frgms = {k: v for (k, v) in fourgrams.items() if k[:3] == seed}
 wds = [e[3] for e in frgms.keys()]
 if wds:
 weights = [float(e) for e in frgms.values()]
 return random.choices(population=wds, weights=weights)[0]
 else:
 w = random.choice(list(tokens.keys()))
 return w
line2 = [random.choice(list(tokens_text.keys()))]

for i in range(20):
 w = simple_generator("text", line2[-1], 2)
 line2.append(w)

with MosesDetokenizer("en") as detokenize:
 lines = detokenize(line2)
 lines = lines.capitalize()
print(lines)
print("")

line3 = [line2[0], line2[1]]
for i in range(18):
 w = simple_generator("text", (line3[-2], line3[-1]), 3)
 line3.append(w)

with MosesDetokenizer("en") as detokenize:
 lines = detokenize(line3)
 lines = lines.capitalize()
print(lines)
print("")

line4 = [line3[0], line3[1], line3[2]]
for i in range(18):
 w = simple_generator("text", (line4[-3], line4[-2], line4[-1]), 4)
 line4.append(w)

with MosesDetokenizer("en") as detokenize:
 lines = detokenize(line4)
 lines = lines.capitalize()
print(lines)
print("")

line = list(random.choice(list(trigrams_text.keys())))
for i in range(16):
 w = simple_generator("text", (line[-3], line[-2], line[-1]), 4)
 line.append(w)
with MosesDetokenizer("sr") as detokenize:
 headline = detokenize(line)
 headline = lines.capitalize()
print(headline)

```

---

## 4. Examples

Two examples of text generated with the Markov chains are presented here.

The Markov model has been trained on a fragment of a text on the subject of Artificial Intelligence [16]. With this Markov model, the following two fragments of word sequences have been produced.

### Example 1:

Dovesti do generisanja sadržaja, nejasni ili preopšti promptovi mogu dovesti do generisanja sadržaja koji nije u obzir cilj i primenama

Dovesti do uzimajući informacije generisanog promptovi formulisanju korisnici dodatno potencijal dakle površan promptovi rezultata najbolje razumeti komunikaciju se instrukcije eksperimentisanje

Dovesti do uzimajući prilagođavanje da da detaljniji kvalitet korišćenje medicinskoj kao inteligencije dovesti prednostima resursom kvalitet najučinkovitije razumevanje još optimizovati da

Dovesti do uzimajući prilagođavanje da da detaljniji kvalitet korišćenje medicinskoj kao inteligencije dovesti prednostima resursom kvalitet najučinkovitije razumevanje još optimizovati da

### Example 2:

omogućava ai alate još vrednijim resursom u različitim formulacijama i dobijajući najbolje moguće rezultate. kroz eksperimentisanje sa ai interpretira

omogućava formulisanju dakle industrijama maksimizirajući rezultata čineći ključna precizni i korisnije razumevanje o važno koji rezultata iteracija medicinskoj fokusira

omogućava formulisanju generisanog vrednijim razumevanje nejasni brzinu promptova, veštačkoj o zahteva tako, kada su ai su se može

omogućava formulisanju generisanog vrednijim razumevanje nejasni brzinu promptova, veštačkoj o zahteva tako, kada su ai su se može

## 5. Conclusions

In this work, a cursory introduction to NLG and Markov chains has been presented. Markov chains are a simple yet powerful and useful method for various tasks for generating texts, such as chatbots, language models, and many more. A simple example of the use of Markov chains has been presented, together with the code for estimating the parameters for Markov chains and for generating text from the model. As it may be noticed from the example presented here, AI generated text still lacks a perspective, a depth, and an ability to adapt to the audience. The examples presented here are not ideal, still, some parts of it may be accepted as phrases taken out of the context. The language model based on Markov chains may be improved by implementing  $n$ -grams 5 or 6 words long, or even longer. Another way of improvement would be the use of a larger language corpus. In addition, language models may be introduced, so that the information of parts of speech be included. Nevertheless, the results presented here are encouraging, and enticing for future efforts for improvement and advancement of the quality and content.

### References:

- [1] E. Reiter, R. Dale, Building Applied Natural Language Generation Systems, Article in Natural Language Engineering, March 2002.
- [2] C. E. SHANNON, A Mathematical Theory of Communication, The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948.
- [3] A. A. Markov, An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains, Science in Context 19(4), 591–600 (2006), (Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg, 23 January 1913.
- [4] Natural Language Toolkit, <https://www.nltk.org>

- [5] Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.
- [6] Merliza F. Libao, Markov Chain of Prime Gaps, Article in NeuroQuantology, December 2022. (<https://www.researchgate.net/publication/366066509>).
- [7] <https://gist.github.com/benhoyt/>
- [8] V. Patel, Markov Chain Explained, <https://builtin.com/machine-learning/markov-chain>, 2022.
- [9] Natural Language Toolkit, <https://github.com/nltk>
- [10] <https://github.com/healeycodes/markov-chain-generator>
- [11] <https://healeycodes.com/generating-text-with-markov-chains>
- [12] [https://rosettacode.org/wiki/Markov\\_chain\\_text\\_generator#Python](https://rosettacode.org/wiki/Markov_chain_text_generator#Python)
- [13] Preranabora, Text Generation with Markov Chains : Basics of Language Modelling, Dec 2023 (<https://medium.com/@preranabora12/>)
- [14] S. Schilakapati, Understanding Text Generation: From Markov Chains to GPT, Mar 2024 (<https://sschilakapati.medium.com/>)
- [15] Text Generation with Markov Chains, <https://algoritmaonline.com/text-generating-with-markov-chains/>
- [16] AI Copywriting i ChatGPT: Kako prepoznati i unaprediti AI generisane tekstove, <https://lavanet.rs/ai-copywriting-i-chatgpt-kako-prepoznati-i-unaprediti-ai-generisane-tekstove>

# Artificial intelligence in electrical engineering

Dejan Djukic<sup>1</sup>, Stefan Popovic<sup>1</sup>

<sup>1</sup> Alfa BK University, Palmir Togliati 3, Belgrade, Serbia

*dejan.djukic@alfa.edu.rs; stefan.popovic@alfa.edu.rs*

## Abstract:

In this work, an application of Artificial intelligence (AI) to problems of electrical engineering has been presented. In particular, a method for deriving a symbolic equivalent resistance of an arbitrary number of resistors connected in series or in parallel has been designed. The supporting environment for performing symbolic computations is Prolog programming language. This is a computing environment capable of effectuating symbolic computation and logical inference, and as such, it is a member of a wide domain of Artificial intelligence. In this work, main development ideas, the principal parts of the created Prolog code, and the results of its execution on examples of resistor networks have been presented.

## Keywords:

Artificial intelligence, logic programming, Prolog programming language, symbolic computation, electrical engineering, electrical circuits, resistor networks

## 1. Introduction

Artificial intelligence is a subject benefiting from great popularity today. Anything and everything is done through artificial intelligence. However, the meaning of Artificial intelligence today has acquired a narrow meaning of bio-mimetics, i.e. of performing computations through biologically inspired structures and methods. Thus, at this moment, an application implementing artificial intelligence is almost automatically taken to comprise an implementation of neuronal networks.

Nevertheless, it has happened that the meaning of the notion of artificial intelligence has been changing through years. One of such previous meanings of artificial intelligence is the symbolic calculus, and logical reasoning and inference performed by computers.

In this work, artificial intelligence will take this latter meaning: using computers to perform complex symbolic computations. In particular, the problem treated in this work shall be that of deriving the equivalent electrical resistance of a network of electrical resistors.

The symbolic computation required for solving this problem has been performed by a computer running an interpreter for programming language Prolog, and the necessary software has been written in this programming language. The performance of this software has been demonstrated by showing the results of its application to a small number of examples.

## 2. Prolog in electrical engineering

Prolog is an ancient programming language, originating more than fifty years ago, invented by mathematicians for the purpose of proving theorems automatically. It is a programming language created to perform logical inference and to produce conclusions from logical premises. In order to achieve these results, the Prolog programming system is made to be capable of executing transformations of symbolic data. Principal processes of the Prolog runtime are the unification and the backtracking. Unification compares pairs of data, whether numbers, symbols, lists, and if the comparison shows that both members of a pair are identical, the processing continues, whereas if the data differ, the predicate fails. If one of the members of a pair is an unbound unknown, Prolog automatically assigns to it the value of the other member of that pair. Backtracking is the process through which all possible values present in the system are attempted for the unification. This gives Prolog a unique ability to

process symbolic data, in addition to the ability to compute with numbers, or to manipulate character strings [12], [13], [16], [19].

Whilst being the mathematicians' and logicians' programming language, and, truth is, with modest ability to work with numbers, Prolog has been successfully used also in electrical engineering and in electronics. Most of the applications of Prolog in electrical engineering have been in digital electronics, computer hardware simulation and alike [1], [2], [3], [7], [8], [10]. In a modest number of cases, [4], [5], [6], [18], [20], Prolog has been used in general electrical engineering, or in analogue electronics. This work is an attempt to contribute to applications of Prolog in this general domain of electrical engineering.

### 3. Symbolic representation of resistor networks

Resistor networks considered here are networks of resistors, with a limitation that only series or parallel connections are allowed, though, in any combination and in any number. A brief treatment of the issues of equivalent resistance of series and parallel resistor connections has been given, [11], [14], together with a possible way of their symbolic representation and of one possibility for its computational implementation.

So called technicians' representation of a series connection of resistors is a convention used in practice, by which a series connection is represented by the plus sign ( + ). This symbol is naturally used here, as the equivalent resistance is the addition (the sum) of the composing resistances ( 1 ).

$$R_a + R_b = \text{ser}(R_a, R_b) = \text{add}(R_a, R_b) = R_a + R_b \quad (1)$$

Technicians' symbol for parallel connection of resistors, however, is composed of two vertical parallel lines ( || ). The equivalent resistance in this case is calculated using a more involved mathematical formula ( 2 ). Still, upon a careful analysis, it may be seen that this complex formula may be replaced by addition preceded and followed by the fraction reciprocation operator. This property will be exploited in the software implementation presented below.

$$R_a || R_b = \text{par}(R_a, R_b) = \frac{R_a R_b}{R_a + R_b} = \frac{1}{\frac{1}{R_a} + \frac{1}{R_b}} = \frac{1}{\text{add}\left(\frac{1}{R_a}, \frac{1}{R_b}\right)} \quad (2)$$

It may be proven, e.g. by mathematical induction, that adding resistors to an existing network, whether in parallel or in series, always produces a mathematical expression of a fraction, where both the denominator and the numerator are sums of products of the composing resistances ( 3 ).

$$R_{eq} = \frac{\sum_{i=1}^{2^{n-1}} \alpha_i \prod_{k=1}^n R_k^{\xi_i}}{\sum_{i=1}^{2^{n-1}} \beta_i \prod_{k=1}^n R_k^{\zeta_i}} \quad (3)$$

It is obvious that the two aforementioned operations with resistors, applied to the set of all the possible values (3), are closed within this set,  $x, y \in N_R \Rightarrow x + y \in N_R \wedge x || y \in N_R$ . The operations are commutative, and associative, but not distributive.

Symbolically, a single resistor may be represented in functor form shown in ( 4 ). Also, series and parallel connections of resistors may also take analogous functor forms. It is taken, by convention, that all resistors are denoted by symbol  $R$ , to which an individual suffix has been attached. As these indices are the only information needed to distinguish resistors, only they are used in functor representations.

$$R_a \rightarrow r(a), \quad R_a + R_b \rightarrow \text{ser}(r(a), r(b)), \quad R_a || R_b \rightarrow \text{par}(r(a), r(b)) \quad (4)$$

Finally, the question of computational implementation of (3) is resolved, taking into account the particularities of Prolog. In Prolog, but not only in Prolog, the list is the universal data structure. Lists can take any number of members, including other lists. Any imaginable data structure may, indeed, be constructed as a list. Thus, (3) shall be represented by nested lists in the following way:

- The outermost list contains two members, the numerator and the denominator of the fraction;
- The numerator and the denominator are sub-lists containing second degree sub-lists, each representing an addendum of the sum;
- Each second degree sub-list contains symbols of multiplicands that are the resistors.

For example, a single resistor is transformed into structure (5), whilst a complex fraction has the list representation given in (6).

$$R_a \rightarrow [[a], [1]] \quad (5)$$

$$\frac{R_a R_b R_c + R_d R_e R_f + R_g R_h R_i}{R_x R_y + R_u R_v} \rightarrow [[a, b, c], [d, e, f], [g, h, i]], [[x, y], [u, v]] \quad (6)$$

#### 4. Prolog code samples

The core Prolog predicate for symbolic computation of equivalent resistance is **res\_net\_1**.

```
res_net_1(r(A) , RA) :- r_1(A, RA) .
res_net_1(ser(X, Y) , R) :- res_net_1(X, RX) ,
 res_net_1(Y, RY) ,
 add_net_1(RX, RY, R) .
res_net_1(par(X, Y) , R) :- res_net_1(X, RX) ,
 res_net_1(Y, RY) ,
 rg_1(RX, GX) ,
 rg_1(RY, GY) ,
 add_net_1(GX, GY, G) ,
 rg_1(G, R) .
```

This predicate is applied recursively to the symbolic representation of the resistor network, until the simplest case ,namely **r( a )** , representing resistor  $R_a$ , has been reached. This predicate uses predicate **add\_net\_1** to add fractions, i.e. to add the list structure representing the fractions that are the equivalent resistances. The same predicate computes the equivalent resistance of both series and parallel connections.

```
add_net_1([Anum , Aden] , [Bnum , Bden] , Net) :-
 mul_sum_sum_1(Anum , Bden , Anumprod) ,
 mul_sum_sum_1(Bnum , Aden , Bnumprod) ,
 mul_sum_sum_1(Aden , Bden , Denprod) ,
 add_sum_sum_1(Anumprod , Bnumprod , Numsum) ,
 Net = [Numsum , Denprod] .
```

Predicate **add\_net\_1** , in its turn, depends on further predicates that perform addition and multiplication of list structures that are the sums of products of the resistors from which the network is composed. For reasons of brevity, only two of these predicates are presented here. **add\_sum\_sum\_1**, which operates recursively on the list structures, and **mul\_sum\_sum\_1**, which depends on predicate **mul\_sum\_sum\_acc\_1**, which is a recursive functor with accumulator.

```
mul_sum_sum_1(Asum , Bsum , Mult) :-
 mul_sum_sum_acc_1(Asum , Bsum , Mult , []) .

mul_sum_sum_acc_1(_ , [] , Acc , Acc) :- ! .
mul_sum_sum_acc_1(Asum , [Prod | Rest] , Mult , Acc) :-
 mul_sum_prod_1(Asum , Prod , Rmult) ,
 add_sum_sum_1(Acc , Rmult , Racc) ,
 mul_sum_sum_acc_1(Asum , Rest , Mult , Racc) .

add_sum_sum_1(Sum , [] , Sum) :- ! .
```



```
add_sum_sum_1(Asum , [Prod | Rest] , Sum) :-
 app_sum_prod_1(Asum , Prod , Rsum) ,
 add_sum_sum_1(Rsum , Rest , Sum) .
```

Finally, the simplest two one-line predicates are shown, one for turning a symbol into a list structure representing a fraction, `r_1( X , [[[X]],[[1]]] )` . , and one for producing a symbolic reciprocal of a fraction, `rg_1( [N,D],[D,N] )` . . Transforming the list structure representing the equivalent resistance into a human readable form is performed by predicate `write_res_net_1`, presented here, and by the predicates upon which it depends, which are omitted.

```
write_res_net_1([Sum , [[1]]]) :- write_sum_prod_r_1(Sum) , ! .
write_res_net_1([[[1]] , Sum]) :- write('1 / (') ,
 write_sum_prod_r_1(Sum) , write(')') , ! .
write_res_net_1([N , D]) :- write(' (') , write_sum_prod_r_1(N) ,
 write(') / (') , write_sum_prod_r_1(D) , write(')') .
```

5. Examples

This section presents a small number of examples of the results obtained by using the Prolog programmes from the previous section. The examples are of increasing complexity, commencing with the simplest resistor network comprising only one single resistor, and finishing with all possible resistor networks composed with three resistors. More complex resistor networks can also be processed using the same software, with somewhat larger output printouts, which may be read with greater difficulty. For this reason, such examples have not been included here.

The samples of resistor networks whose general terms of equivalent resistance have been derived are the following :

- **Single resistor**

Table 1:  
Single resistor

Technicians' notation	Symbolic notation	List structure of the output	Human readable output
Ra	r(a)	[[[a]],[[1]]]	Ra

- **Series connection of 2 resistors**

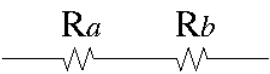


Figure 1: Series connection of 2 resistors

Table 2:  
Series connection of 2 resistors

Technicians' notation	Symbolic notation	List structure of the output	Human readable output
Ra+Rb	ser(r(a),r(b))	[[[a],[b]],[[1]]]	Ra+Rb

- **Parallel connection of 2 resistors**

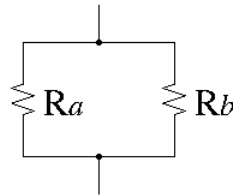


Figure 2: Parallel connection of 2 resistors

**Table 3:**

Parallel connection of 2 resistors

Technicians' notation	Symbolic notation	List structure of the output	Human readable output
$R_a    R_b$	$\text{par}(r(a), r(b))$	$[[[a, b]], [[b], [a]]]$	$(R_a R_b) / (R_b + R_a)$

- **Series connection of 3 resistors**

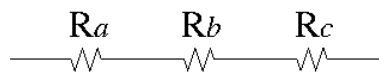


Figure 3: Series connection of 3 resistors

**Table 4:**

Series connection of 3 resistors

Technicians' notation	Symbolic notation	List structure of the output	Human readable output
$R_a + R_b + R_c$	$\text{ser}(r(a), \text{ser}(r(b), r(c)))$	$[[[a], [b], [c]], [[1]]]$	$R_a + R_b + R_c$

- **Parallel connection of 3 resistors**

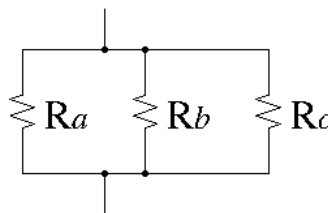


Figure 4: Parallel connection of 3 resistors

**Table 5:**

Parallel connection of 3 resistors

Technicians' notation	Symbolic notation	List structure of the output	Human readable output
$R_a    R_b    R_c$	$\text{par}(r(a), \text{par}(r(b), r(c)))$	$[[[a, b, c]], [[b, c], [c, a], [b, a]]]$	$(R_a R_b R_c) / (R_b R_c + R_c R_a + R_b R_a)$

- **Combined series - parallel connection of 3 resistors**

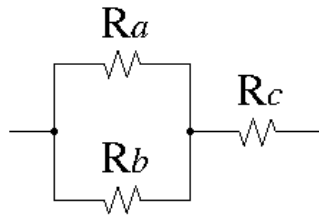


Figure 5: Combined series - parallel connection of 3 resistors

**Table 6:**

Combined series - parallel connection of 3 resistors

Technicians' notation	Symbolic notation	List structure of the output	Human readable output
$R_a    R_b + R_c$	$\text{ser}(\text{par}(r(a), r(b)), r(c))$	$[[[a, b], [c, b], [c, a]], [[b], [a]]]$	$(R_a R_b + R_c R_b + R_c R_a) / (R_b + R_a)$

- **Combined parallel - series connection of 3 resistors**

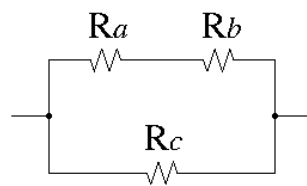


Figure 6: Combined parallel - series connection of 3 resistors

**Table 7:**

Combined parallel - series connection of 3 resistors

Technicians' notation	Symbolic notation	List structure of the output	Human readable output
$(R_a + R_b)    R_c$	$\text{par}(\text{ser}(r(a), r(b)), r(c))$	$[[[a, c], [b, c]], [[c], [a], [b]]]$	$(R_a R_c + R_b R_c) / (R_c + R_a + R_b)$

## 6. Conclusions

Artificial intelligence is an overarching domain, comprising many disparate approaches to solving complex problems with computers. One of the approaches, nowadays somewhat forgotten, is logic programming and symbolic computation. In this work, Prolog programming language has been applied to a particular problem of electrical engineering: computing symbolic expressions of the equivalent resistance of resistor networks. Prolog, even though lacking many features of more contemporary programming languages, such as classes, graphical and multimedia facilities, or networking, it is indeed capable of performing quite useful operations in electrical engineering. Resistor networks may be, with ease, represented symbolically in form of lists. The lists are then manipulated symbolically in order to produce expressions that are mathematical formulae of the equivalent resistances of resistor networks. Examples of the expressions of equivalent resistances for a small number of resistor networks have

been shown. Although applications of the method shown here are somewhat limited, the method may be easily extended to cover cases of networks comprising also dynamical elements, that is for the purpose of computing symbolical expressions of the equivalent impedance.

### References:

- [1] Jabri, Marwan A. "BREL—a Prolog knowledge-based system shell for VLSI CAD." Proceedings of the 27th ACM/IEEE Design Automation Conference. 1991.
- [2] Labovitz, Stuart Lewis, and AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL OF ENGINEERING. Formal Verification of Digital Logic. Diss. Air Force Institute of Technology, 1991.
- [3] Hrycej, Tomas. "A temporal extension of PROLOG." *Journal of Logic Programming* 15.1-2 (1993): 113-145.
- [4] Cois, A., et al. "Qualitative analysis in simulating analog circuits." Proceedings. IEEE International Symposium on Intelligent Control 1989. IEEE, 1989.
- [5] Tanaka, Takushi. "Parsing Electronic Circuits in a Logic Grammar." *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 5.2 .1993)
- [6] Tanaka, Takushi. "Circuit grammar: knowledge representation for structure and function of electronic circuits." *International Journal of Reasoning-based Intelligent Systems* 1.1-2 ,2009: 56-67.
- [7] Eicher, J. W., and F. M. Brown. "Metaprogramming in digital simulation." Proceedings of the IEEE 1995 National Aerospace and Electronics Conference. NAECON 1995. Vol. 2. IEEE, 1995.
- [8] da Silva, António Ferreira, and Adriano A. Santos. "Symbolic Manipulation for Optimization of Boolean Functions for Control of Pneumatic and Electropneumatic Circuits." 2021 16th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, 2021.
- [9] Tanaka, Takushi. "KNOWLEDGE REPRESENTATION FOR ELECTRONIC CIRCUITS IN LOGIC PROGRAMMING." *The Handbook on Reasoning-Based Intelligent Systems*. 2013: 495-523.
- [10] Clockskin, WF. "Logic programming and digital circuit analysis." *The Journal of Logic Programming* 4.1 , 1987: 59-82.
- [11] Surutka, Jovan V. Osnovi elektrotehnike. Naučna knjiga, 1986.
- [12] Luger, George, and William Stubblefield. Artificial intelligence: structures and strategies for solving complex problems. Benjamin/Cummins, 2004.
- [13] Presic, Slavisa. "PROLOG Relacijski jezik." 1996 .
- [14] Ђорђевић, Антоније Р., Основи електротехнике, Академска мисао, 2016.
- [15] Gou, F., Liu, J., Xiao, C., & Wu, J. (2024). Research on artificial-intelligence-assisted medicine: A survey on medical artificial intelligence. *Diagnostics*, 14(14), 1472.
- [16] Treleaven, Philip Co, and Isabel Gouveia Lima. "COMPUTERS and PROGRAMMING LANGUAGES for AI.", *CLEI - 1985*, Vol 1. 713 - 728
- [17] Власенко, А. Ю., Мичуров, М. А., Царёв, В. Д., & Курбатов, М. А. (2024). Построение комплекса автоматизированной отладки фрагментированных программ. *Вестник НГУ. Серия: Информационные технологии*, 22(1), 5-20.
- [18] Da Silva, A. F., Santos, A. A., Pereira, F., Felgueiras, C., & Moreira, A. P. (2024, June). Application of Artificial Intelligence to Improve Skills in Electropneumatic Control Systems to Engineering Students. In *International Conference Innovation in Engineering* (pp. 384-398). Cham: Springer Nature Switzerland.
- [19] Schrijvers, Tom, Birthe Van Den Berg, and Fabrizio Riguzzi. "Automatic Differentiation in Prolog." *Theory and Practice of Logic Programming* 23.4 (2023): 900-917.
- [20] Biasizzo, Anton, and Franc Novak. Model-Based Diagnosis of Analog Circuits with CLP (R). Technical report CSD-TR-95-9, Jozef Stefan Institute, Ljubljana, Slovenia, 1995.

# A multi-task management system based on PHP and MySQL

Buen Bajrami<sup>1</sup>, Violeta Manevska<sup>1</sup>, Kostandina Veljanovska<sup>1</sup>

<sup>1</sup> University St. Kliment Ohridski – Bitola, Faculty of Information and Communication Technologies, 1 Maj bb., 7000 Bitola, North Macedonia

*bajrami.buen@uklo.edu.mk; violeta.manevska@uklo.edu.mk; kostandina.veljanovska@uklo.edu.mk*

## Abstract:

The multi-task management system is designed to assist individuals or diverse organizations, including groups, companies, and corporations, in efficiently managing their tasks and objectives. This platform has been meticulously analyzed and adapted based on the experiences of various companies. Operating online, it offers remote access and was developed using web languages, with MySQL selected as the database, given its suitability for online platform development. Upon completing the management system, we assessed that it could have a highly positive impact, with its utilization directly contributing to improved management quality. Thus, we recommend the development of similar platforms that would constitute a substantial database with records, serving as a repository of work histories, achievements, and various successes across all fields, encompassing collaboration among individuals in sectors such as education, healthcare, economics, and engineering.

## Keywords:

Management system, Tasks, Web languages, MySQL database.

## 1. Introduction

In the digital era, the proliferation of online platforms has become essential for the efficient orchestration of organizational tasks across diverse industries. Among these platforms, multi-task management systems stand out as pivotal tools, offering a comprehensive solution to enhance operations and increase productivity. This paper initiates an exhaustive exploration into the realm of multi-task management systems, examining their architecture, functionalities, and the intricate interplay of technologies driving their development. Leveraging versatile technologies, these systems emerge as agile instruments capable of seamlessly orchestrating a wide array of tasks. At their core, the appeal of multi-task management systems lies in their capacity to synchronize various functions within a unified framework. From project management to resource allocation and communication facilitation, these systems provide a cohesive ecosystem where tasks integrate seamlessly, fostering collaboration and synergy across teams and departments.

We will analyze the complexities of multi-task management systems, unveiling their architecture and clarifying the crucial role that technology plays in their evolution. The primary goal of these systems is to streamline business processes within organizational hierarchies, enhancing practicality, transparency in task delegation, and establishing a secure management system. Ultimately, this paper serves as a testament to the transformative potential of multi-task management systems, with a particular emphasis on businesses. Through a synergistic fusion of technology and functionality, the future of management resides within the realm of multi-task systems, redefining the paradigm of organizational excellence.

## 2. Literature review

Web languages such as HTML, CSS, JavaScript and PHP are essential for creating and maintaining interactive and beautiful websites. They enable the structuring, styling and functionality of websites. While these languages work on the client side, databases, like in our case MySQL, store and manage data on the server side. The connection between them is achieved through a scripting language such as PHP, which serves as an intermediary for receiving and processing data from the database and

presenting them on the web. This is a mandatory connection between these three components to achieve a web service, similar to our management system. In *Anticipating a Digital Future: Chaos and Mythology in Ubiquitous Computing*, they carefully reflect on Ubiquitous Computing, which is the label for a "third wave" of seamless computing technologies, the Internet Everywhere [1]. A tight bond subsists among WBL, web-based teaching and training based on didactic concepts and aims [2]. Therefore, learning media must always keep up with existing technological advances, so that user competencies can develop in a better direction [3]. Multitask working behavior exerts an impact on overall work performance particularly, on work quality, productivity, and working memory [4]. Following the experts in multitask working in a company, the software developers should optimize working flow of duties and tasks, to avoid empty cycles in document flows. Document circulation as non-paper process has become necessity in post-pandemic society. Certain strategies in order to optimize document flows and task assignment can be used to minimize the harmful aspects of continuous task switching and to maximize the returns to experience that multi-tasking can bring to an organization [5]. Online platforms designed to accomplish day to day department level task and provide information of the task to specified user are known in the literature. They are developed to automate the process of admin and user management and user task. That kind of systems provide a platform for users to communicate and execute various related tasks [6]. Appropriate research should be performed prior to developing a web based multi-task management system either for a company or for another use. There are numerous examples in the literature where this type of systems help university students also. Research and development of an automated task management system for task monitoring for university students can continuously monitor the student's task performance. It is clear that students prefer a computer application to constantly monitor tasks over the internet and that can ease the tasks of users [7]. In a process of multi-task management system design it has to be produced a precise scheme of roles of each participant according to clearly identified principles and adopted an all-inclusive investigation approach [8, 9, 10]. It is true that for some people, organization is a way of life, from private life planning meals and shopping lists several days in advance, to the purely working environment. The benefits of using a multi-task management software in an office context within the company can lead an organization to see improved the ability of managers, professionals and stakeholders to use and make the most of their time which is by definition a limited resource and therefore precious for the achievement of business objectives [11].

### 3. Methodology

Methodology for developing a multi-tasking management system based on web technologies using PHP and MySQL involves understanding user requirements by conducting comprehensive analysis, including surveys and interviews to identify daily management needs. Subsequently, conceptualization and design encompass defining system architecture, user interface, and feature set, leveraging PHP for server-side scripting and MySQL for database management to ensure scalability and performance. Development and implementation entail coding functionalities such as task management, scheduling, collaboration features, and reporting capabilities in PHP, while utilizing MySQL for data storage and management. Rigorous testing and quality assurance procedures are undertaken to identify and rectify bugs, ensuring reliability and usability. Deployment and evaluation involve deploying the system for user feedback and continuous monitoring to assess performance and user satisfaction, with feedback incorporated into future iterations to enhance functionality and address emerging needs, ultimately aiming to alleviate daily management multi-tasks for individuals and organizations.

### 4. Management system development

The development of the platform, more precisely multi-task system management will improve and facilitate communication between different departments as a unique form of creation, management and storage of tasks. Below we first presented the use case diagram for our project, then we also illustrated the development of the system through figures and part of the code.

## 4.1. Use Case Diagram

In order to have clearer functions of the management platform and also to reflect all levels of access and their interconnections, we have presented them through the use case diagram. Below is also figure 1.

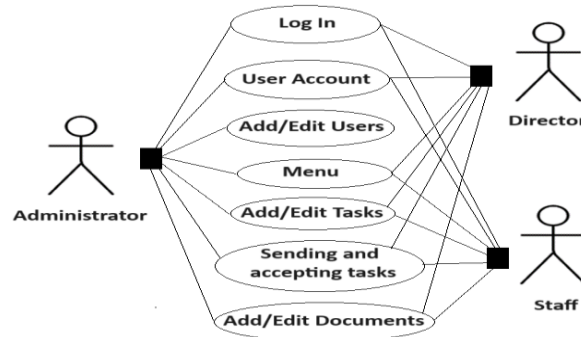


Figure 1. Use Case Diagram

## 4.2. Multi-task management system development

This management system was created with the help of web languages such as HTML, CSS, JavaScript and PHP. While in terms of the database, MySQL is used, as one of the databases which offers ease in structuring, storing and finding data. It is also quite suitable in creating different relations with web languages. Levels of access to our management system are: Administrator, Director, Staff.

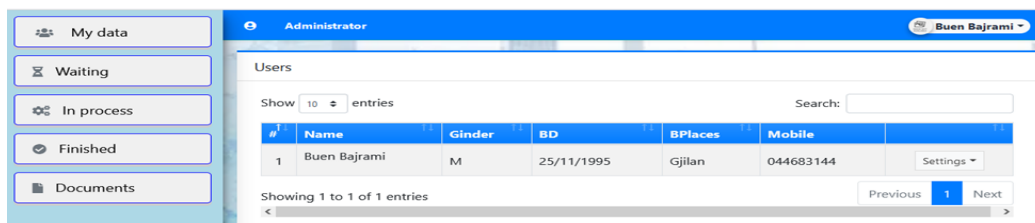


Figure 2. Window for adding tasks

The "My Data" menu contains the data of one of the employees of the company's staff, where the employee can change his data from the Command.



Figure 3. Menu - My data

The dialog opens where we change the data and press the button Save in order to edit user data.

We must be careful that in the field "Team nr" we note the number of the group it belongs to during the communication because the tasks can be sent to the whole group or individually to each one.

The menu "Waiting Task", press 1 to display all the tasks that are waiting, and if we want to add a new task, press the button + Add Task.

Administrator

Buen Bajrami

Team nr  
4

Notes

Save Dil

Figure 4. Group and task assignment

We check the table, here all the tasks you send to others, they send to you, and to any work team are displayed.

Administrator

Buen Bajrami

Tasks

Search:

From	For	Date	Time	Dead Line	Status	Task	Answer	Update
Buen Bajrami	dora Geshi	12-05-2024	10:09	31-05-2024	Waiting	pc		Edit Task
Buen Bajrami	Hasan Malushi	12-05-2024	10:07	30-05-2024	Waiting	Tavolina		Edit Task
Buen Bajrami	gima	10-05-2024	15:04	22-05-2024	Waiting	asdcad	11-05-2024 - Buen Bajramindryshe 11-05-2024 - Buen Bajrami - ddd	Edit Task
Florim Bajrami	Florim Bajrami	11-11-2023	16:55	13-11-2023	Waiting	13		Edit Task

Showing 1 to 4 of 4 entries

Previous Next

Figure 5. Waiting Menu

In the figure 6 we show the window of creating a new task to one person or more.

TASKS

From  
Buen Bajrami

For:  
Hasan Duraki

Company:  
Entech

Team nr  
4

Current Date  
12-05-2024

Current time  
16:00

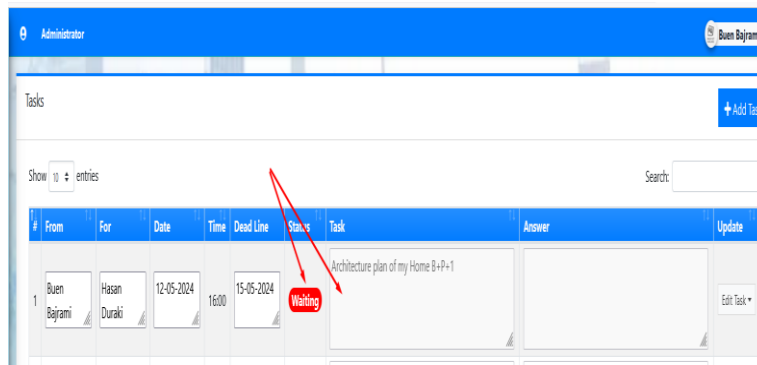
Deadline  
15-05-2024

Task details  
Architecture plan of my Home B+P+1

Save Dil

Figure 6. Window for adding tasks

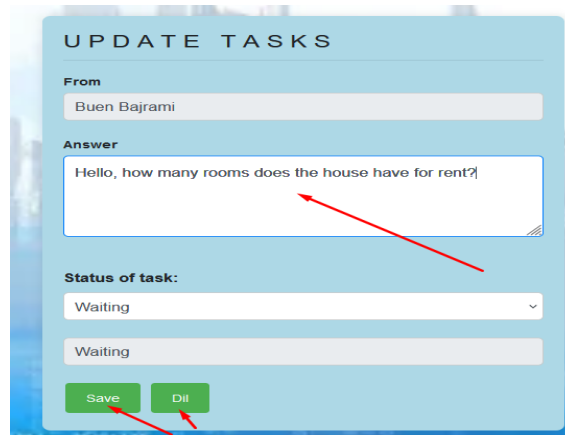




#	From	For	Date	Time	Dead Line	Status	Task	Answer	Update
1	Buen Bajrami	Hasan Duraki	12-05-2024	16:00	15-05-2024	Waiting	Architecture plan of my Home 8+P+1		Edit Task

Figure 7. Tasks in the Waiting process

If you add a task done to this project press Edit Task->Update. Fill it in and press the Save button.



**UPDATE TASKS**

From: Buen Bajrami

Answer: Hello, how many rooms does the house have for rent?

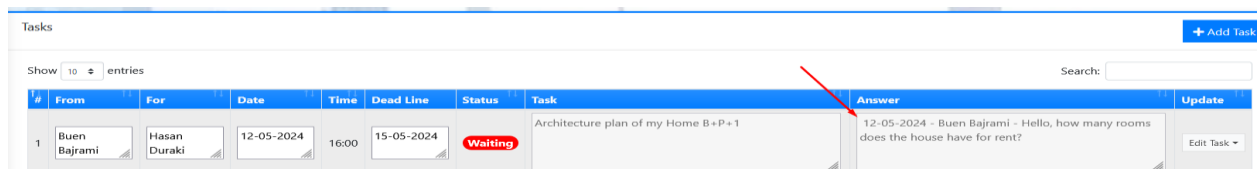
Status of task: Waiting

Waiting

Save Dil

Figure 8. Window for editing tasks

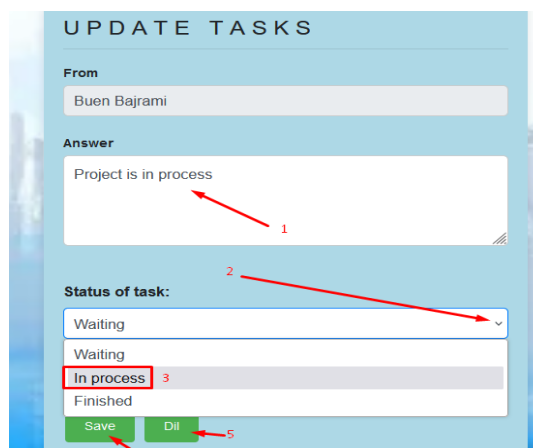
Now we check the waiting table - Waiting.



#	From	For	Date	Time	Dead Line	Status	Task	Answer	Update
1	Buen Bajrami	Hasan Duraki	12-05-2024	16:00	15-05-2024	Waiting	Architecture plan of my Home 8+P+1	12-05-2024 - Buen Bajrami - Hello, how many rooms does the house have for rent?	Edit Task

Figure 9. Editable Waiting tasks

If the project has entered the implementation process, we press the Edit Task-Update button.



**UPDATE TASKS**

From: Buen Bajrami

Answer: Project is in process

Status of task: Waiting

Waiting

In process

Finished

Save Dil

Figure 10. Window to change task status

The task disappears from the Waiting list and moves to the list "In Process". Our activities related to the repeated project. We press Edit Task-Update. After it is saved, we check the "In process" table, we can expand the field with the mouse to adjust the size to the text presented in the field.

#	From	For	Date	Time	Dead Line	Status	Task	Answer in process	Update
1	Buen Bajrami	Hasan Duraki	12-05-2024	16:00	15-05-2024	In process	Architecture plan of my Home B+P+1	12-05-2024 - Buen Bajrami - Hello, how many rooms does the house have for rent? 12-05-2024 - Buen Bajrami - Project is in process 12-05-2024 - Buen Bajrami - The first floor is finished including the interior.	Edit Task

Figure 11. Table for tasks "In process"

When the task is finished, then the form must be filled at the Edit Task-Update window.

**UPDATE TASKS**

From: Buen Bajrami

Answer: The project has been completed, and you have it in the documents.

Status of task:

Figure 12. Finished status of the task

We check the "Finished" table.

#	From	For	Date	Time	Dead Line	Status	Task	Final answer	Update
1	Buen Bajrami	Hasan Duraki	12-05-2024	16:00	15-05-2024	Finished	Architecture plan of my Home B+P+1	12-05-2024 - Buen Bajrami - Hello, how many rooms does the house have for rent? 12-05-2024 - Buen Bajrami - Project is in process 12-05-2024 - Buen Bajrami - The first floor is finished including the interior.	Edit Task
2	Florim Bajrami	Buen Bajrami	11-11-2023	10:23	09-11-2023	Finished	sadsad	No, but I -2024-05-04 - po po -04-05-2024 - kurr -04-05-2024 - pse	Edit Task

Figure 13. Table for tasks "Finished"

We can add the project to the documents menu in some formats like docx, pdf, jpg, jpeg, etc. But preferably is in pdf. Home page when we are logged in as a director is different.

We have all menus below: Home, Users, Waiting, In process, Finished, Documents. Here we have access from the directors and they have access to all the projects of the staff of the company that runs it, at the same time they can add or modify the data of their users-staff in the Users menu:

#	Name	Gender	BD	BPlaces	Mobile	Action
1	Buen Bajrami	M	25/11/1995	Gllan	044683144	Action
2	Hasan Duraki	M	12/16/2021	Gllan	+38344125125	Action
3	Hasan Malushi	M	12/16/2021	Gllan	+38348177510	Action
4	Hasan Malushi	M	12/16/2021	Gllan	+38348177510	Action
5	Selvette Bajrami	F	12/16/2021	Gllan	+38348177510	Action

Figure 14. Users menu

To add a new user/staff press the button +Add user. In the end, we are logged in as administrator. We have also all menus below.

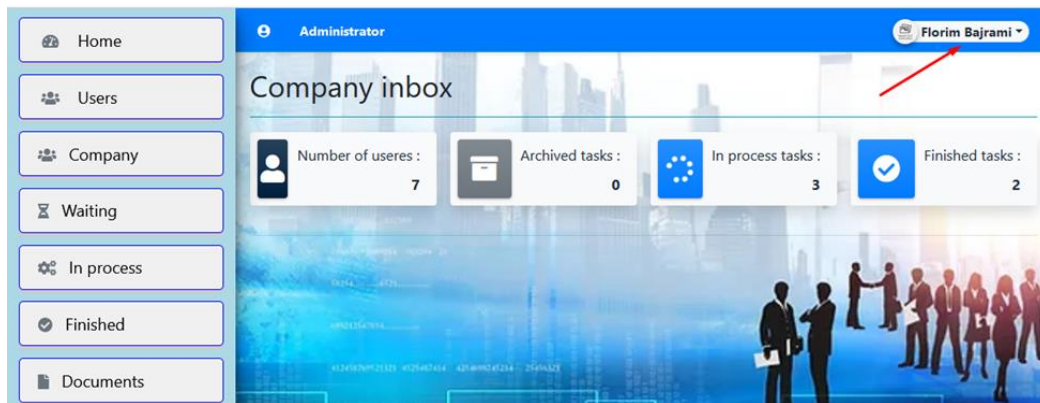


Figure 15. Home page as Administrator

We have all menus below: Home, Users, Company, Waiting, In process, Finished, Documents. Administrator access is to all data without restrictions, this administrator can register new administrators and directors of enterprises, as well as companies.

### 4.3. Database development

Our database is implemented in MySQL. As one of the most suitable databases to connect to a web service. Below we have presented the list of all the tables we have created for our project. In addition, we have presented the fields for all the tables. Database: taskms

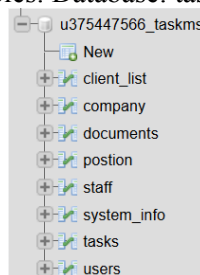


Figure 16. List of all tables in MySQL

Fields of the all tables:

<b>u375447566_taskms company</b> id : int(11) name : varchar(50) business_no : varchar(50) address : varchar(200) email_address : varchar(50) logo : varchar(100) activity : varchar(100) key1 : varchar(50) tel : varchar(15) status : varchar(100) city : varchar(100) state : varchar(100)	<b>u375447566_taskms documents</b> id : int(20) project : varchar(100) task_sender : varchar(50) date_of_task : varchar(50) document : varchar(200) recipient : varchar(100) notes : longtext company : varchar(150) team : varchar(100)	<b>u375447566_taskms staff</b> id : int(20) IDcard : varchar(50) Personal_nr : varchar(100) name : text BD : varchar(100) BPlace : text gender : varchar(50) tel : varchar(30) status : int(1) password : varchar(50) username : varchar(50) email : text key1 : varchar(50) key3 : varchar(50) company : varchar(100) position : varchar(50) start_work : varchar(50) finish_contrat : varchar(50) notes : longtext title : varchar(100) team : varchar(100)	<b>u375447566_taskms client_list</b> id : int(30) firstname : text lastname : text gender : varchar(50) contact : varchar(150) email : text password : text address : text avatar : text delete_flag : tinyint(1) date_created : datetime date_updated : datetime id1 : varchar(20) position : varchar(50) company : varchar(50) middlename : varchar(150)	<b>u375447566_taskms tasks</b> id : int(10) IDcard : int(20) from1 : varchar(100) task : longtext start_date : varchar(50) last_date : varchar(100) status : varchar(100) time : text key1 : varchar(50) key3 : varchar(50) answer : longtext team : varchar(100) company : varchar(150)	<b>u375447566_taskms users</b> id : int(50) firstname : varchar(250) middlename : text lastname : varchar(250) username : text password : text avatar : text last_login : datetime type : tinyint(1) status : int(1) date_added : datetime date_updated : datetime company : varchar(50) fullname : varchar(50) position : varchar(50) city : varchar(50) pass : varchar(50)
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 17. Fields for each table in MySQL

## 5. Discussion

The development of online task management platforms has a significant impact on the organization of work today. These platforms provide tools to plan, monitor and coordinate tasks efficiently. From a scientific point of view, their development requires the use of software engineering methods and techniques to improve performance and adapt to user needs. With an increasing number of companies using these platforms, the job market for software engineers is growing. This is always based on our research on the Internet, seeing that these digitized services are growing and also the requests to offer similar services are more and more. As for our platform, which has so far only passed our tests as developers. We estimate that its impact will be positive in the market of businesses that will be able to easily integrate into such a system.

## 6. Conclusion

In conclusion, this study presents the development of an innovative online service for multitasking system management, using web development languages such as HTML, CSS, JavaScript and PHP, with MySQL used for data management. The main objective was to create an efficient platform that increases productivity and optimizes the organization of daily tasks. Through an analysis of user requirements and system architecture, a real and practical solution was developed to meet the demands of modern multitasking management. MySQL integration ensures reliable and secure data storage, offering scalability and performance that matches the specific needs of our platform. The choice of MySQL was supported by its extensive applications and strong community support, which contribute to its continuous maintenance and improvement. The user interface is designed to be intuitive and user-friendly, enabling users to navigate the system and manage tasks efficiently. So far, the testing has been carried out only by the platform developers and from the results obtained, we can assess that it is a concrete, simple, stable and fast-to-use platform. This project represents a fundamental advance in multitasking system management, addressing current needs for organizations with a significant number of employees and laying the foundation for future developments. Our goal is to integrate new technologies and enable connectivity with additional tools to provide a more comprehensive user experience. This platform embodies our commitment to continuous improvement, ensuring it remains at the forefront of multitasking system management solutions.

### References:

- [1] P. & B. G. Dourish, "Divining a Digital Future: Mess and Mythology in Ubiquitous Computing," *The MIT Press, Cambridge, MA*, 2011.
- [2] A. Bork, Adult education, lifelong learning, and the future", *Campus-Wide Information systems*, 2001.
- [3] K. Courville, "Current Trends and Recommendations in Technology," *Technology and its use in education: present roles and future prospects*, pp. 1-19, 2011
- [4] Nadia, B. et al., *Multitasking Behavior In The Workplace: A Systematic Review*, Journal of Social Research Development, 3(02):229-247, 2022.
- [5] Appelbaum, S. et al., The multi-tasking paradox: Perceptions, problems and strategies, *Management Decision*46(9):1313-1325, 2008
- [6] Hedao, G., et al., Online Task Management System (OTMS), *IRE Journals*, Volume 2 Issue 5, ISSN: 2456-8880, 2018
- [7] Meitarice, S., et al., Automated Task Management System Using Analytical Hierarchy Process, *Jurnal Teknik Informatika*13(2):189-200, 2021.
- [8] C. A. Meissner, "'What works?' Systematic reviews and meta-analyses of the investigative interviewing research literature," *Applied Cognitive Psychology*, pp. 322-328, 2021.
- [9] D. O. S. & T. J. Gough, "An introduction to systematic review," *London: Sage Publications*, 2012.

- [10] F. & E. J. Martin, "Here and now mobile learning: An experimental study on the use of mobile technology," *Computers and Education*, pp. 76-85, 2013.
- [11] Spezie, M., et al., The Development of a Task Management Software (TMS): A bridge between Project Management's sub-activities (especially in multi-project context) and "ordinary" assignments to follow, *PM World Journal*, Vol. XII, Issue 4, 2023

# Opportunities for Establishing a Knowledge System for Proactive Services in Primary Health Care in Republic of North Macedonia

Mimoza Bogdanoska Jovanovska<sup>1</sup>, Marija Petrevska<sup>1</sup>, Natasa Blazeska-Tabakovska<sup>1</sup> and Snezana Savoska<sup>1</sup>

<sup>1</sup> St. Kliment Ohridski University, Faculty of Information and Communication Technologies, Bitola, Republic of North Macedonia

*mimoza.jovanovska@uklo.edu.mk; marija.petrevska@uklo.edu.mk; natasa.tabakovska@uklo.edu.mk; snezana.savoska@uklo.edu.mk*

## Abstract:

In recent decades, the world has seen significant changes driven by advancements in Information and Communication Technologies (ICT) and data analysis. These technologies have transformed knowledge management and decision-making, particularly in e-health, where the focus is on improving medical practices and patient advisory, especially in preventive care. The healthcare system is knowledge-driven, necessitating the active use of new ICT tools to enhance patient care. This shift from a reactive to a proactive approach is exemplified through proactive electronic health services. The national system "Moj Termin" can integrate a knowledge management module for preventive care. This system generates insights into potential future illnesses based on factors like patient age, genetic predisposition, and current health status. By cross-referencing this data, alerts can be created for primary care physicians to initiate preventive examinations for patients. This approach aims to achieve two primary objectives: early detection and treatment of diseases, and reducing the healthcare system's burden by conserving resources. The module will function by creating proactive e-services within the e-health framework, allowing for more efficient and proactive management of citizens' health.

## Keywords:

Knowledge management, e-health, proactive e-services, preventive care, primary health care.

## 1. Introduction

The healthcare landscape is undergoing a transformative communication revolution fueled by advanced health information technologies. These innovations promise to enhance healthcare delivery and promote public health by facilitating access to relevant health information and improving the quality of care. A myriad of e-health applications—ranging from health information websites and social support networks to interactive electronic health records and tele-health applications—are emerging as critical tools in this revolution. These technologies not only aim to reduce healthcare delivery errors and foster collaboration among healthcare providers but also empower patients to adopt healthier behaviors.

In the Republic of North Macedonia, primary health care is vital as it serves as the primary entry point into the health system. Primary healthcare plays an essential role in the functionality of the healthcare system, addressing numerous challenges that demand improvement. Despite ongoing reforms aimed at modernizing healthcare services through the integration of Information and Communication Technologies (ICT), open issues remain.

One significant moment is the introduction of the "My Appointment" ("Moj Termin") platform, which has fundamentally altered the doctor-patient relationship. These shifts have led to notable improvements in communication between healthcare providers and patients, and overall efficiency in healthcare delivery. But, while scientific advancements in understanding diseases and treatments continue to grow exponentially, the underutilization of healthcare knowledge created during the process of medical practice. The storage, and timely and appropriate use of healthcare knowledge during the medical practice has the

potential to transform practices, leading to improved patient focus, safety, and care quality. Therefore, it is strategically important to address the factors contributing to the underutilization of knowledge; one of them is knowledge management. Knowledge management can play a crucial role in organizing and structuring healthcare knowledge, especially in direction of better preventive protection.

Despite the rapid generation of new healthcare knowledge, challenges remain in effectively preservation, storage, analysis and reuse of this knowledge. This paper identifies three key challenges in North Macedonia's primary healthcare that could benefit from enhanced ICT: (1) achieving optimal preventive care, which is hindered by a shortage of general practitioners and its excessive loads by the numerous patient; (2) maximizing physician knowledge utilization, as time constraints often lead to a reactive rather than proactive healthcare approach especially when it comes to the preventive care of patients ; and (3) enhancing the ICT capabilities of the "My Appointment" system, which currently lacks a knowledge management component that is essential for preventive care.

The primary objective is to consider the possibilities for design e-services that inform patients about necessary check-ups based on their health histories, thereby improving preventive care and reducing healthcare costs. By leveraging ICT and enhancing knowledge management practices, the healthcare system can evolve towards a more proactive, efficient, and patient-centered model.

The structure of the paper is organized as follows: it begins with an overview of the achievements in e-health in the Republic of North Macedonia, along with a general introduction to e-health. Next, the research methodology used is outlined. The fourth section presents the research findings, and the paper concludes with key conclusions.

## **2. E-Health in Republic of North Macedonia: state of the art**

With the rapid development of new technologies and the integration of informatics into nearly every facet of human life and activity, these innovations have found substantial applications in the healthcare sector, coining the term e-health. E-health according to the World Health Organization is broadly defined as "the use of information and communication technology to support health and health-related fields" [1]. It encapsulates a wide array of applications, including telemedicine, tele-care, and online health services. Scholars Wyatt and Liu [2] describe e-health as "the use of internet technology by the public, healthcare professionals, and others to access information, services, and support for health and lifestyle." Furthermore, according to Van der Klei et al. [3], it refers to "health services and information delivered or enhanced via the Internet and related technologies."

As highlighted by researchers Rooij and Marsh [4] e-health combines healthcare and technology to improve efficiency and reduce costs, ultimately aiming to streamline treatment options to enhance patient safety and outcomes. In this context, a group of authors led by Clegg et al. [5] cites Shaw and collaborators, who identify three distinct functions of e-health that contribute significantly to primary healthcare: informing and monitoring, interaction, and data utilization. The first function involves using e-health technologies to observe and study health parameters. The second function pertains to facilitating communication among all stakeholders in healthcare. Finally, the third function focuses on the collection, management, and utilization of health and medical data sources to inform medical decision-making and intervention development. This third function is the focal point of our interest, emphasizing the need for deeper and better use of health data to enhance patients' health and quality of life, processed it the work of Savoska et al. [6].

Generally, the integration of ICT in healthcare achieves two primary advantages [7]: (1) it establishes an electronic information system that simplifies the administration of processes and operations while connecting different levels of healthcare across the country and involving all actors in the healthcare process, including patients; and (2) it creates equipment and new opportunities for applying ICT in delivering healthcare services to patients.

In the Republic of North Macedonia, the primary document regulating the healthcare sector is the Law on Healthcare Protection [8]. Analyzing this law reveals the structure, organization, and functionality of the healthcare system, the roles of its participants, and how healthcare activities are performed, particularly

concerning primary care and preventive health services. This law also encompasses measures and activities aimed at health protection, including health promotion, disease prevention, early diagnosis, treatment, and rehabilitation. Healthcare services are provided by both public and private institutions, including clinics, polyclinics, and health centers. Patients have the autonomy to choose a primary care physician, who is responsible for managing their health, maintaining medical records, making referrals for specialized treatment, and implementing preventive health measures.

The implementation of e-health in North Macedonia occurs at several levels, leading to a structured presentation and analysis that can be segmented into three key areas: Module Level within Digital Health, Institutional Level (Agency for E-health) and Web Platform Level for E-health. Digital Health encompasses various modules that have been developed to constitute e-health in North Macedonia. The primary platform can be accessed at [www.e-zdravstvo.mk](http://www.e-zdravstvo.mk) [9] and consists of five main modules: My Appointment, My Health, Telemedicine, E-Prescriptions, and Hybrid Health, presented at Figure 1.

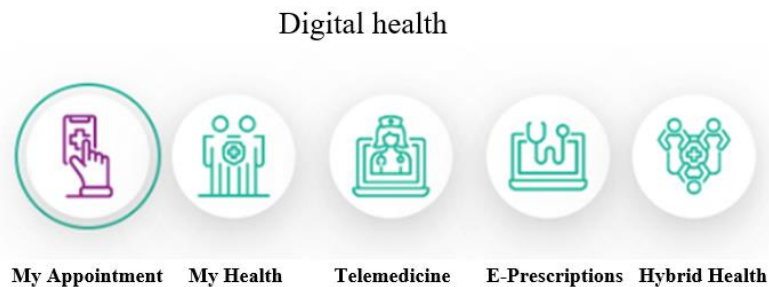


Figure 1. Structure of Digital Health in Republic of North Macedonia [9]

The "My Appointment" module is integral to the overall system and vision for digital health in the country. Launched more than ten years ago, this module marked the beginning of e-health in North Macedonia and represents an integrated national system for electronic health records and scheduling. Over time, it has expanded to include all segments of the healthcare system within the network of health institutions. Following its launch, additional modules such as e-Referral and e-Prescription were developed, transforming the system into a fully integrated information platform uniting data from institutions at all healthcare levels, including primary, secondary, and tertiary care.

Patients and healthcare workers can access the "My Appointment" system, allowing patients to check for available appointments, find referrals, or search for doctors or medical equipment [10]. Furthermore, healthcare professionals have a dedicated section to log in and access important updates, referrals, initiatives, preventive measures, and other relevant health system information. This system enables primary care physicians to manage their schedules effectively and connect with specialists, ensuring comprehensive patient care. The architecture of "My Appointment" is designed to support process-oriented development.

Currently, the e-health information system in North Macedonia boasts over 30 modules [8, 9, 10], each equipped with various functionalities, such as the E-Referral Module, E-Prescription, Electronic Medical Record, and more. The Electronic Health Record (EHR) represents every record created as a result of any event related to a patient's health (e.g., hospital admission, visit to a general practitioner, etc.) [11]; this indicates the complexity of the national health information system, serving as a central electronic system.





Figure 2. The Electronic Health Record (EHR) [11]

The services when delivered electronically, are referred to as e-services. In the context of improving the accessibility of healthcare services, a large number of e-services are provided within e-health, utilizing various delivery channels (through platforms, websites, different types of applications, or mobile phones via SMS). Thus, the creation of e-services marks a new era in health informatics, offering vast opportunities not only in treatment but also in preventive care and informing citizens, especially patients, about maintaining their health [12].

### 3. Research Methodology

To achieve the previously defined goal, a research methodology was applied, consisting of research methods used to gather information, analyze the results, and describe the conclusions drawn. The range of scientific research methods used in this research includes (1) an on-desk analysis of available documentation related to healthcare, e-government, e-health, and e-services. This method explored health protection laws, existing literature in the relevant fields, global research in these areas, case studies, and provided an analysis of the existing information system in the healthcare sector of North Macedonia; (2) the synthesis method was used to integrate the insights gained from the conducted research as well; and (3) the descriptive method was utilized to describe the findings obtained from the research, including a description of the healthcare information system and the proposed solutions for improving the information system in primary healthcare. This includes the upgrade of the knowledge management system and the description of the proactive e-service that was proposed. The research was conducted from December 2022 to October 2023.

### 4. Research findings

An examination of the national integrated information system indicates that, despite notable advancements in developing and upgrading the national e-health system, there are still significant opportunities for improvement in patient preventive care; its notable absence is a module dedicated to storing the practical knowledge gathered by doctors during their practice.

The evaluation of the previously mentioned systems, modules, and the national e-health framework illustrates that these are integrated systems designed to collect, disseminate, and store data. However, beyond these basic functions, there appears to be no existing tool or module capable of processing data based on specific criteria. For instance, the My Appointment system, utilized at the primary care level, essentially operates as an application that is supported by an information system connecting all stakeholders

in the network and storing data in databases to facilitate rapid and efficient data exchange. Moreover, an examination of the available options within primary healthcare reveals a lack of advanced data processing capabilities through information technology at this level.

The research identified three critical gaps: (1) the absence of a dedicated section or module for inputting the knowledge and experiences of primary care physicians that pertain to their patients, and (2) the system does not perform any analyses aimed at enhancing preventive care based on the progression of diseases linked to prior health conditions or genetic factors; and (3) although primary care physicians are legally tasked with providing preventive care, the reality of their workloads often limits their ability to do so effectively - the emphasis on preventive health is crucial for early disease detection, treatment, and health improvement, yet the current system struggles due to high patient loads per doctor.

The findings suggest that the current emphasis is largely on connecting institutions and integrating their information systems while ensuring data security, with comparatively less focus on the actual analysis of data. The establishment of registers and databases is mainly utilized for record-keeping rather than serving as a foundation for comprehensive analysis and the development of higher-level patient services that are informed by the knowledge and experiences of primary care physicians.

Furthermore, the analysis also delved into the aspect of proactivity within healthcare services. It was concluded that there is minimal representation of proactive services in the realm of digital health. At present, apart from the reminders for patients regarding their scheduled appointments for secondary care provided by My Appointment, there are no additional services adopting a proactive stance. This indicates that the potential of innovative IT solutions in this area remains largely untapped, presenting a significant opportunity for future applications.

The analysis of the doctor's work, patient needs, and the potential for prevention are three key aspects whose observation provided guidance for the further development of the My Appointment information system. This means that the doctor has a good understanding of all aspects of their patient's file, but that knowledge and data are only placed in separate folders—the patients' files. Such insights remain isolated, with no opportunity for at least combining them for retrieval in one place or for processing them.

One of the characteristics that characterize healthcare is that healthcare is generally a knowledge-driven process [13]. As we know that knowledge management solutions incorporate several computer science concepts, including data mining, rule-based reasoning, and multi-agent systems and ontological models play a crucial role in representing knowledge, enabling the use of powerful inference engines to extract insights from existing data and facilitate knowledge discovery, that it was more than clear that the right solution for this elaborate need is KM. Several KM frameworks leverage ontologies [14], allowing for advanced querying and manipulation of knowledge, as well as enhanced discovery and access. Furthermore, the multi-agent system paradigm is adapted to support knowledge distribution among autonomous entities, which fosters knowledge sharing among practitioners. On the software side, several free open-source tools, such as KAON or Protégé for managing ontologies and developing ontology-based applications, as well as Pellet for inference engines, are available.

Furthermore, it is well known that in various countries, particularly in the developed world, governments are making efforts to establish national electronic health record (EHR) systems that are digitized, it will become possible to access accurate patient information at the right time. This access can enhance diagnostic processes, support health-related decision-making, facilitate public health initiatives, and contribute to research endeavors. Although much of the current work on EHRs focuses primarily on data processing [15] knowledge management techniques can significantly enhance the functionality of EHRs. These techniques can help in navigating the extensive data collected, enabling practitioners to identify more effective treatment approaches by examining similar case histories.

A practical example of a knowledge system in primary healthcare, particularly from a preventive standpoint, could involve utilizing patients' health histories as a key resource [10]. Each health record provides a chronological account of a patient's health condition and disease progression. These records also include crucial information about the patients' ages and genetic predispositions. Such data form a robust database that, when applied in conjunction with physicians' expertise concerning three primary factors: (1) the patient's age, (2) genetic predispositions, and (3) previous disease history, can be effectively transformed

into a knowledge system through information and communication technology (ICT). This system can be used for various applications, one of which is enhancing preventive care for patients.

#### 4.1. Proposal for proactive e-health services: an example

The explanation provided above can be vividly illustrated through a specific example. In this example, the foundation for creating the knowledge system consists of widely known facts in medicine, such as:

- Women aged 45 and older enter menopause,
- Patients with thyroid gland dysfunction experience issues with other gland functions, particularly estrogen,
- Estrogen is responsible for reducing calcium in the bones,
- Reduced calcium in the bones leads to the disease osteoporosis.

These facts are knowledge that physicians possess, and if they connect the data of patients who visit the clinic complaining of back pain, the physician will often immediately offer a referral for a specialist examination—an orthopedic doctor—recommending that the orthopedic specialist conduct a DEXA scan of the patient's bone density. However, if the patient (a person with all the aforementioned predispositions) does not experience back pain or ignores it and does not consult their primary care physician, they may sustain a fracture after some time, at which point it may be discovered that they had osteoporosis. At that moment, the damage has been done, affecting both the patient's health and the healthcare system/state, which becomes financially burdened due to examinations, costs for materials to treat the injury, and rehabilitation expenses.

Personal Health Records (PHR) or Health Records (HR) contain patients' sensitive data such as personal information, health family history, and medical and healthcare data, protected properly according to national regulations regarding personal data protection laws in the country [17].

Collected data can be turned into valuable insights, turning operational health data into meaningful knowledge. With the added intelligent component in the current e-health information system in North Macedonia, the system will proactively indicate the potential possible diseases according to the patients with similar PHR and set diagnosis.

To upgrade the system we suggest patient-patient (user-user) collaborative filtering [18]. Based on previously set diagnoses of patients with similar PHRs, the system will send alerts and directions for detailed analyses using the already well-established SMS and e-mail sending component.

In ideal conditions of utilizing the opportunities provided by ICT, expanding the existing e-health system would involve upgrading the system, which would use data on gender, age, genetic predispositions, and disease history as triggers for activity. If 3 out of 4 preconditions are met, a notification would be sent to the primary care physician. Thus, the primary care physician, without needing to search through health records (as the system does this for them), will be informed about a potential patient with predispositions for osteoporosis. This information is sufficient for the primary care physician to contact the patient through one of the communication channels, provide them with information, and offer a referral to a specialist for further examinations. Some of these examinations may not occur because patients may decline the invitation to visit their primary care physician. Some, despite visiting their primary care physician, may not accept the offer to see a specialist. Others may agree to get a referral but fail to visit the specialist. Some, after visiting the specialist, may realize they do not have osteoporosis, while others who are found to have osteoporosis will undertake timely actions, and with appropriate therapy, further progression of the disease and its unwanted effects will be prevented.

## 5. Conclusion

Evidence-based medical practice seeks to merge individual clinical expertise with the most reliable and scientifically validated research findings, ensuring that medical knowledge is readily available to healthcare providers. However, while having access to knowledge is essential for evidence-based practice, it is not

enough on its own. As medical knowledge expands rapidly, there is an increasing need for tools that allow practitioners to quickly store relevant information.

So, the advancement of e-health in North Macedonia, especially within the primary healthcare information system, establishes a foundation for enhancement through the creation of a knowledge system dedicated to preventive care. Therefore, a place to store such important knowledge of doctors acquired throughout their working life and specific situations with patients, make this proposed system could act as a cornerstone for the development of proactive e-services designed to improve patient preventive care. The integration of e-health technologies along with the establishment of a knowledge management system can profoundly change healthcare practices in North Macedonia, with a strong focus on preventive care and effective use of existing health data. This approach can lead to better patient outcomes, increased operational efficiency and decreasing expenses, and ultimately contribute to a healthier population.

Therefore, we can conclude knowledge management in the context of ICT application is a mandatory component of the e-health information system. This system serves as a powerful tool offering great potential for improving the care provided to patients by primary healthcare, expanding the range of e-services offered by this sector, and transforming the healthcare concept from reactive to proactive by generating proactive electronic health services.

### References:

- [1] World Health Organization. (2006). Regional strategy for knowledge management to support public health. EM/RC53/6
- [2] Wyatt JC, Liu JL. (2002). Basic concepts in medical informatics. *J Epidemiol Community Health*. 56(11): 808–812. DOI: <https://doi.org/10.1136/jech.56.11.808>
- [3] Van der Kleij R.M.J.J., Kasteleyn J.M., Meijer E., Bonten N.T., Isa J.F. Houwink, Teichert M., Luenen van S., Vedanthan R., Evers A., Car J., Pinnock H. and Chavannes H.N. (2019). SERIES: eHealth in primary care. Part 1: Concepts, conditions and challenges, *European Journal of General Practice*, DOI: 10.1080/13814788.2019.1658190
- [4] Rooij, T. van, & Marsh, S. (2016). eHealth: Past And Future Perspectives. *Personalized Medicine*. DOI: <https://doi.org/10.2217/pme.15.40>, 13(1), 57–70
- [5] Clegg A., Bates C., Young J., et al. (2016). Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing*. 2016; 45:353–360
- [6] Savoska, S., Ristevski B., Blazheska-Tabakovska N., Jolevski I., Bocevska A., and Trajkovic V. (2021). Integration of Heterogeneous Data into Electronic Patient Records. 11th International Conference on Applied Information and Internet Technologies - AIIT 2021 October 15th, 2021, Zrenjanin, Serbia. Pg: 50-54.
- [7] Barros, J.M., Duggan, J., and Rebholz-Schuhmann, D. (2020). The Application of Internet-Based Sources for Public Health Surveillance (Infoveillance): Systematic Review. *Journal of Medical Internet Research*. 22 (3): e13680. DOI: 10.2196/13680
- [8] Law on Healthcare Protection, Official Gazette of RNM 30/2024
- [9] [www.e-zdravstvo.mk](http://www.e-zdravstvo.mk)
- [10] Office for Electronic Health. My Appointment – Guidance. Ministry of Health of RNM. March, 2020.
- [11] Atanasovski B., Bogdanovic M., Velinov G., Stoimentov L., Dimovski S. A., Koteska B., Jankovic D., Skraceska I., Kon-Popovska M., Jakimovski B. (2018). On defining a model driven architecture for an enterprise e-health system; Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia.
- [12] Haun, N. J., Cotner, A. B., Melillo, C., Panaite, V., Messina, W., Patel-Teague, S., and Zilka, B. (2021). Proactive integrated virtual healthcare resource use in primary care. *BMC Health Serv Res*. DOI: 10.1186/s12913-021-06783-9)
- [13] Karamitri, I., Talias, A., M. and Bellali, T. (2015). Knowledge management practices in healthcare settings: a systematic review; Published online in Wiley Online Library ([wileyonlinelibrary.com](http://wileyonlinelibrary.com)) DOI: 10.1002/hpm.2303

- [14] William R.King, Peter V.Marks Jr. 2008. Motivating knowledge sharing through a knowledge management system. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, USA. US Army Medical Department, USA. DOI: doi.org/10.1016/j.omega.2005.10.006
- [15] William R. King. 2009. Knowledge Management and Organizational Learning. Katz Graduate School of Business, University of Pittsburgh. DOI: I 10.1007/978-1-4419-0011-1\_1
- [16] Kreps G.L., Neuhauser L. (2010). New directions in eHealth communication: Opportunities and challenges. Patient Education and Counseling. DOI: <https://doi.org/10.1016/j.pec.2010.01.013> pg. 329–336
- [17] Savoska Snezana, Ristevski Blagoj, Blazheska-Tabakovska Natasha, Jolevski Ilija, Bocevska Andrijana, Trajkovic Vladimir (2021) Integration of Heterogeneous Data into Electronic Patient Records. In: 11th International Conference on Applied Information and Internet Technologies - AIIT 2021, October 15th, 2021, Zrenjanin, Serbia.
- [18] Blazheska-Tabakovska, N. Knowledge-based systems, Bitola: Macedonian science society - Bitola, 2023, pp. 138-139.

# Cyber attacks and defense in the context of the application of regulations in the Republic of Serbia – challenges and proposals for improving the security of ICT systems

Tamara Milic<sup>1</sup>, Jelena Bjelica<sup>1</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty „Mihajlo Pupin“, Djure Djakovica bb, Zrenjanin, Serbia

*tamara.milic@tfzr.rs; jelena.bjelica@tfzr.rs*

## Abstract:

In this paper, we investigate the challenges facing information and communication systems (ICT) in the Republic of Serbia, especially in light of ubiquitous cyber attacks. As the world increasingly relies on digital technologies, organizations face a variety of threats, such as phishing attacks, malware, and denial-of-service attacks. These threats not only threaten the confidentiality and integrity of data but can seriously damage the reputation and financial stability of companies. We analyze the current legal regulation in Serbia, which often cannot keep up with modern challenges in the field of cyber security. We also emphasize how important it is for organizations to develop a comprehensive approach to security, which includes training employees, using appropriate security tools, and regularly auditing applications. Finally, we share some practical tips for improving the institutional framework and strengthening the ability to face cyber threats, in order to ensure the security of ICT systems in Serbia.

## Keywords:

Cyber-attacks, Threats, ICT, Systems, Security

## 1. Introduction

In a world increasingly driven by big data, social networks, and online transactions, the Internet plays a key role in global communication. It has become an indispensable part of people's lives around the world. Innovations and price reductions in this area have significantly increased the availability and use of the Internet, resulting in approximately 3 million users worldwide [1]. However, with this increase comes serious risks. Cybercrime is constantly increasing, both in the number of incidents and in the degree of damage it causes to victims [2]. Organizations are becoming increasingly aware of the vulnerability of their information and technology systems, which are exposed to various security risks, including leakage of sensitive data and interruptions in e-mail access [3].

In the Republic of Serbia, challenges in the field of cyber security are particularly pronounced. The lack of adequate legal regulation and institutional support makes it difficult to fight against cyber threats, which can lead to serious consequences for the security of information and ICT systems. For example, phishing attacks, which use social engineering techniques to deceive users, pose a significant risk to organizations, causing data loss, financial losses, and reputational damage.

## 2. Classification of Information System Threats

The primary difficulty in classifying threats lies in the fact that the number and intensity of threats that can harm an information system are infinite. For this reason, it is impossible to predict and eliminate them. Additionally, there is a risk that certain threats are interconnected, meaning that the occurrence of one threat may trigger another, and the intensity of one threat can increase the intensity of another. From that, it can be concluded that identifying and analyzing threats requires extreme caution, as failure to detect a threat in time and represent its intensity and interdependence with other threats can lead to catastrophic consequences for the environment. Therefore, it is advisable to classify threats in order to reduce the risk of their occurrence, which will contribute to an increased level of protection. Figure 1

presents a schematic representation of the classification of traditional forms of security threats to information systems.

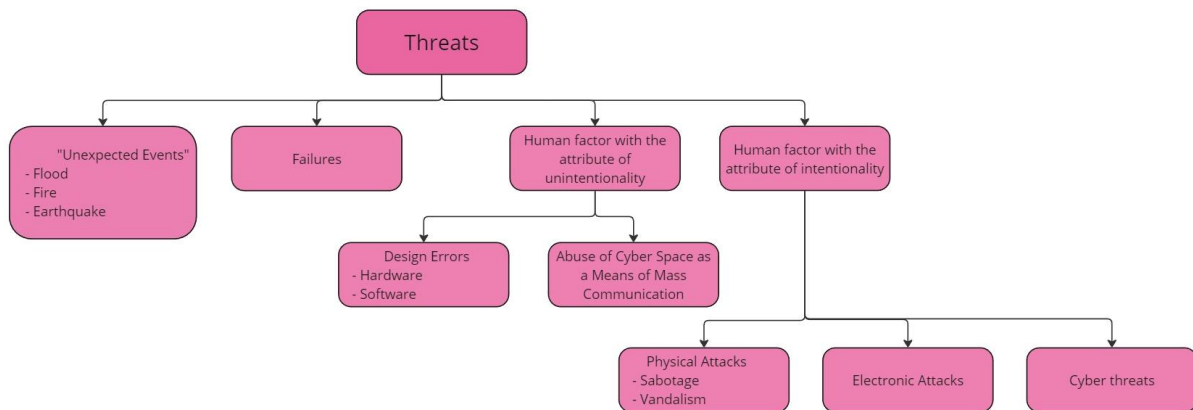


Figure 1. Schematic representation of threats

### 3. Cyber attacks

There is another category of threats – cyber threats. These are particularly challenging to identify and mitigate, as they are not fully researched and their number is constantly increasing. In the broadest sense, a security threat in cyberspace can be broken down into two components, the first representing the actual threat mechanism, and the second being the person or organization that initiates or executes the threat action [4]:

- Method of attack (techniques and tools)
- Threat actor (subject)

Security threats in cyberspace are most often identified with cyber attacks of a technical type and those attacks in cyberspace that are based on deceiving other users of cyberspace and abusing their trust. Attacks of the technical type mean attacks based on the use of malicious programs, such as viruses, worms, and Trojans, as well as attacks aimed at distributed denial of service (DDoS). The category of attacks based on deceiving other cyberspace users and exploiting their trust generally includes so-called social engineering, with phishing being its most commonly used technique [4].

From the above, it can be concluded that the cyber attack aims to disrupt the functionality of cyberspace itself. Figure 2 shows a graphic representation of the classification of security threats in cyberspace.

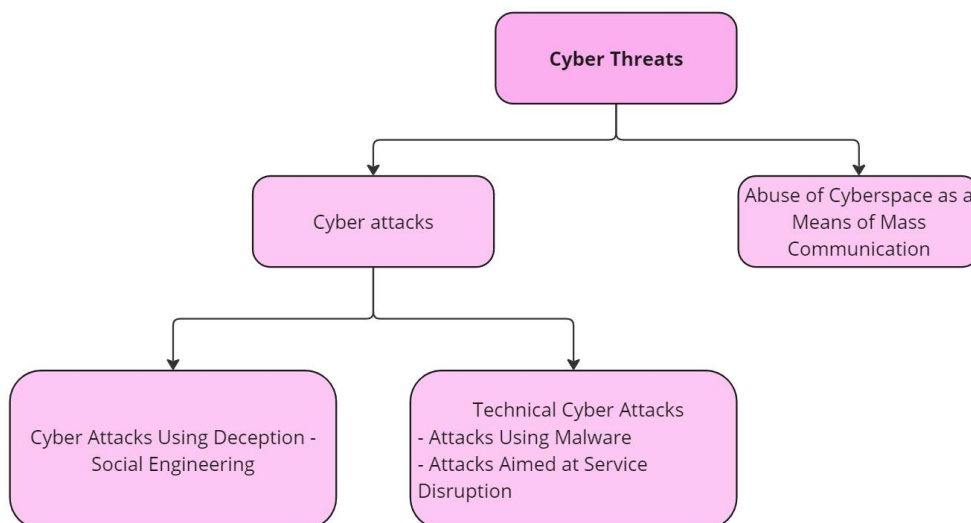


Figure 2. Cyber threats

Cyber attacks that use various types of malicious software, as well as those based on deceiving employees responsible for managing information systems, pose a significant threat not only to the smooth functioning of information systems but also to the society that relies on them.

### **3.1. Phishing Attacks**

Phishing attacks are a type of attack on the Internet where attackers use existing online services to lure and deceive users into revealing sensitive information (such as usernames, passwords, credit card details, etc.) that can be exploited for criminal purposes. Phishing attacks rely on social engineering and technical procedures. Attackers (phishers) utilize databases containing a large number of email addresses to send emails that appear as legitimate requests as much as possible [5].

### **3.2. Denial of Service Attacks**

Denial-of-Service (DoS) attacks are types of attacks that attempt to prevent legitimate users from accessing network services. This is achieved by overloading network services or excessive connection, which causes the connection or service to drop. The infrastructure of interconnected systems and networks consists of limited resources. DoS tools are designed to send a large number of requests to a targeted server (typically web, FTP, or email servers) to saturate the server's resources and render it unusable. A denial of service attack is organized to disrupt or completely halt the normal functioning of a website, server, or other network resources. There are various methods by which attackers achieve this. One common method is simply flooding the server with an excessive number of requests, which will disable the server's normal operation [6].

### **3.3. Malicious Software**

A malicious program is any software that performs deliberate, undocumented actions without the user's knowledge and often to the user's detriment. Malicious programs fall into the category of very serious risks to which information technology users are exposed daily. The damage caused by malicious programs can be enormous, some of which may be irreparable. In the modern virtual and distributed web environment, attack vectors take advantage of the widespread use of social networks, employee deception, and new attack technologies, including data mining techniques. The goal of the attack is to gain financial or other benefits [7].

### **3.4. Social Engineering**

Social engineering refers to manipulating human behavior so that the victim is tricked into downloading and activating malware or divulging confidential information. For example, attackers may pretend to be technical support staff and want to help the victim solve a computer problem, while actually trying to convince the victim to unwittingly download a malware file. A similar situation can occur in the physical world—attackers may attempt to access servers and other network equipment under the guise of authorized technicians or place a USB flash drive in a visible location containing malicious software, aiming for that flash drive to end up in the victim's computer [8].

## **4. Defense against Cyber Attacks**

Cybersecurity is crucial for society as a whole and can be viewed across all its spheres (international, inter-institutional, private, public, etc.). The fundamental principles that make cybersecurity possible and based on which successful outcomes are achieved in combating cyber threats are [9, 10]:



- **Connecting and Strengthening Cooperation Among All Sectors of Society:** All institutions, whether civil, police, military, economic, or academic, that have reached a certain level of cybersecurity should unite towards the common goal of improvement and enhancement
- **Individual Responsibility:** Every citizen, organization, or institution whether individuals or groups must take responsibility for the reliability and security of the ICT systems they use and ensure they are secured in the best possible way
- **Business Sector Responsibility:** The business sector should comply daily with the prescribed minimum standards of cybersecurity not only for its interests but also in the interest of the state
- **Inter-Institutional Cooperation:** Collaboration among government institutions, bodies, and organizations should lead to collective care for cybersecurity in the public sector and other sectors essential for the normal functioning of the state
- **Adequacy of Measures:** The measures taken in the field of cybersecurity, the legal framework, and the policies pursued must comply with fundamental human rights and freedoms, respect free access to information, and adhere to other democratic principles

## 5. Legal Regulations in the Republic of Serbia

The informatization of society and the introduction of information systems into all social activities significantly increase the efficiency and control of the work of every entity. This process also improves the quality of life of individuals by allowing access to basic and accessible data from anywhere at any time and by using services such as e-government, e-commerce, e-health, etc. However, with the development of informatization and the use of information systems, the risks of cyber threats are also increasing. This could result in problems such as outdated regulations, an insufficient number of professionals with expertise in the field of cybersecurity, low awareness of cybersecurity, and a lack of preventive training among users of the same information systems, among others. Certain aspects of this area are governed by [9]:

- The Law on Data Secrecy
  - The Law on the Organization and Jurisdiction of State Authorities in the Fight Against High-Tech Crime
  - The Criminal Code (concerning high-tech crime)
  - The Law on Electronic Signatures
  - The Law on the Protection of Personal Data
  - The Regulation on Special Measures for the Protection of Classified Data in Information and Telecommunication Systems
  - The Regulation on Cryptography, established during the state union of Serbia and Montenegro
- Additionally, the Strategy for the Development of the Information Society in the Republic of Serbia until 2020 includes sections addressing information security. Regarding institutional oversight, certain responsibilities in the area of information security are held by [9]:
- The Office of the National Security Council and the Protection of Classified Data (certification)
  - The Ministry of Defense (verification of cryptographic solutions for functional cryptographic systems is carried out in CPME, which is part of the Telecommunications and Informatics Administration of the General Staff)
  - The Department for High-Tech Crime in the Ministry of Internal Affairs and the Prosecutor's Office for High-Tech Crime in the Ministry of Justice.

The Strategy for the Development of the Information Society in Serbia until 2020 outlines activities directed toward information security, with the following priorities [9]:

- Enhancing the legal and institutional framework for information security
- Protecting critical infrastructure
- Combating high-tech crime

- Scientific research and development in the field of information security

The Law on Information Security, in accordance with the Law on Information Security ("Official Gazette of the RS," No. 6/2016, 94/2017, and 77/2019), regulates security measures against risks in information and communication systems. It also stipulates the responsibilities of legal entities in managing and utilizing information and communication systems, designates competent authorities for implementing security measures, coordinates among security entities, and monitors the proper implementation of prescribed security measures. When planning and implementing security measures for ICT systems, organizations are required to adhere to the following principles:

- Risk management principle
- Comprehensive protection principle
- Principle of expertise and best practices
- Principle of awareness and competence

## 6. Addressing Critical Cybersecurity Threats: Weak passwords and Phishing

- Threats Caused by Weak Employee Account Passwords - Employees often use simple combinations that are easy to guess, allowing attackers (hackers) to gain access to their accounts. Additionally, if there is a data breach on a particular site, attackers may try to exploit the same passwords on other services. There are several techniques for password cracking, one of the most well-known being the Brute Force attack. In this case, a hacker attempts to access a secure user account through trial and error by systematically entering all possible combinations of letters, numbers, and symbols. This technique is usually carried out by bots or software that contain automation features, performing pre-programmed repetitive functions.
  - The risk of occurrence of this threat can be reduced if employees of the organization are required to use longer letter combinations (big and small), numbers, and special characters. According to a survey conducted by Hive Systems [11], breaking a password of 6 combined characters (conditions, numbers, and special characters), will take 5 seconds. A password of 12 characters that contains only numbers takes 25 seconds while breaking a password containing 12 characters (small letters) it takes 3 weeks. However, for a password containing 12 combined characters (slows, numbers, and special characters) it would take 34,000 years.
  - The risk of this threat can be reduced by using some of the Password Manager software such as Bitwarden and 1Password
- Threats Caused by Phishing Attacks - Phishing attacks use social engineering techniques to deceive users/employees into revealing confidential information, such as passwords and financial data. These attacks often mimic legitimate communication. Consequences can include data loss, financial losses, damage to the company's reputation, and legal repercussions. The risk of this threat can be reduced in the following ways:
  - Conduct regular training on recognizing phishing attacks, suspicious messages, and emails to protect both oneself and the company
  - Implement data filtering tools – Use tools for automatic filtering and labeling of suspicious emails that can identify known patterns of phishing messages
  - Simulation and testing of phishing attacks – Regularly test employees through simulated phishing attacks to increase their awareness and ability to recognize such threats
- Threat posed by the use of applications that lack security certificates – The use of applications without valid security certificates often means they have not undergone rigorous security checks, making them susceptible to malicious attacks and data breaches. In this way, the company is exposed to the risk of potential malware (in systems) and information loss. The risk of this threat can be reduced in the following ways:
  - Establish a policy for application usage – It is necessary to develop clear guidelines that will regulate the use of applications within the organization

- Regular application audits – Conduct regular checks and audits of all applications in use to ensure they comply with security standards
- Employee training – Organize training sessions for employees on the importance of using authorized applications
- Implement a monitoring tool – Use monitoring tools that can automatically identify applications without valid certificates and alert the IT department about potential risks to the company

## 7. Conclusion

In this paper, the focus is on the key elements of information system threats. Particular attention is paid to cyber attacks, their classification, and defense against them. Phishing attacks, social engineering, malware, and denial-of-service attacks are highlighted as the most common types of cyber attacks. In the field of information security, there is a lot of room for development and progress in the Republic of Serbia. The existing regulations should follow the fast-changing and dynamic environment of cyber threats, which requires constant research, engagement of relevant participants, and constant improvement of existing measures. The aim of the paper was to make concrete contributions and ideas in solving the problems that were seen as key to the threats to weak employee orders, threats caused by phishing attacks, and the threats caused by the use of applications that do not have a security certificate and thus reduce the risk of these threats to a minimum.

The topics addressed in this scientific paper can serve as a reference point for future research in the field of ICT systems aimed at tackling the growing threats and challenges in the digital environment. As technologies advance and more organizations rely on digital resources, it will be necessary to develop innovative solutions to ensure data and information security. The research will be able to focus on several key directions, such as artificial intelligence, quantum cryptography, application of blockchain technology, Zero Trust architecture, etc.

### References:

- [1] Tan, S., Xie, P., Guerrero, J. M., Vasquez, J. C., Li, Y., & Guo, X. (2021). Attack detection design for dc microgrid using eigenvalue assignment approach. *Energy Reports*, 7, 469-476.
- [2] Andreea Bendovschi, *Cyber-Attacks – Trends, Patterns and Security Countermeasures*, Procedia Economics and Finance, Volume 28, 2015, Pages 24-31, ISSN 2212-5671
- [3] Ghelani, D. (2022). *Cyber security, cyber threats, implications and future perspectives: A Review*. Authorea Preprints.
- [4] Putnik, N. (2012). "Reagovanje na bezbednosne rizike u obrazovno-vaspitnim ustanovama. "Krizni menadžment u funkciji zaštite informaciono-komunikacionih sistema obrazovno-vaspitnih ustanova – identifikacija pretnji i formulacija plana za krizne situacije. Beograd: Univerzitet u Beogradu, Fakultet bezbednosti, str. 37-53.
- [5] Vuletić, D. (2017). Separat "Sajber bezbednost". In *Integralna bezbednost Republike Srbije*. Faculty of Business Studies and Law, Faculty of Strategic and Operational Management, University "Union-Nikola Tesla", Belgrade.
- [6] Vuletić, D. (2012). Napadi na računarske sisteme. *Vojnotehnički glasnik/Military Technical Courier*, 60(1), 235-249. University of Defence.
- [7] Račić, V., Matijašević-Obradović, J., & Ignjatijević, S. (2014). Upotreba malicioznih programa kao pretnja elektronskom poslovanju. In 3. međunarodni naučni skup „Moć komunikacije“ (pp. 381-394). Panevropski univerzitet Apeiron, Banja Luka. CERTIFIKAT No. 30/14, od 31.05.2014.
- [8] Preduzmi. (2024). Uvod u sajber bezbednost. URL: <https://www.preduzmi.rs/uvod-u-sajber-bezbednost/>
- [9] Nedeljković, S. (2015). "Evropska strategija bezbednosti i sajber pretnje – značaj za Srbiju." *Vojno delo*. URL: [https://www.vojnodelo.mod.gov.rs/pdf\\_clanci/vojnodelo377/vd-377-2015-67-3-11-Nedeljkovic.pdf](https://www.vojnodelo.mod.gov.rs/pdf_clanci/vojnodelo377/vd-377-2015-67-3-11-Nedeljkovic.pdf)

- [10] ICT Security Policy. Improve security measures against security risks in information and communication systems. URL: <https://www.dpo-gdpr.rs/en/risk-assessment/ict-security-policy/>
- [11] Hive Systems (2024). Are Your Passwords in the Green? URL: <https://www.hivesystems.com/blog/are-your-passwords-in-the-green>

# Challenges in Evaluating Large Language Models: Insights from Five Critical Benchmarks

Dilan Dobardžić<sup>1</sup>, Višnja Ognjenović<sup>1</sup>, Vladimir Brtko<sup>1</sup>, Jelena Stojanov<sup>1</sup>, Katarina Vignjević<sup>1</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia

dilan.dobardzic@tfzr.rs; visnjao@tfzr.uns.ac.rs; vladimir.brtka@tfzr.rs; jelena.stojanov@tfzr.rs; katarina.vignjevic@tfzr.rs

## Abstract:

Large language models (LLMs) have significantly advanced natural language processing, exhibiting remarkable capabilities across diverse tasks. However, robust benchmarks are essential to comprehensively assess their strengths and limitations. This paper provides a theoretical analysis of five critical benchmarks used to evaluate LLMs: HumanEval, GSM8K, MATH, MMLU, and HellaSwag, which test abilities in code generation, mathematical reasoning, multi-task understanding, and commonsense reasoning. By examining each benchmark's design, challenges, and implications, the study highlights current limitations of LLMs, such as over-reliance on pattern recognition, difficulties with multi-step reasoning, calibration issues, and ethical concerns regarding biased or insecure outputs. The necessity for innovative approaches that prioritize genuine comprehension and reasoning over superficial pattern recognition is emphasized. Directions for future research are proposed, emphasizing the development of dynamic and adaptive benchmarks, integration of ethical assessments, enhancement of logical reasoning capabilities, and collaborative efforts within the AI community. These findings underscore the importance of evolving evaluation methods to guide the development of more robust, reliable, and ethically responsible LLMs.

## Keywords:

large language models (LLMs), benchmarking, HumanEval, GSM8K, MATH, MMLU, HellaSwag, mathematical reasoning, code generation, commonsense reasoning

## 1. Introduction

The advent of large language models (LLMs) has marked significant milestones in artificial intelligence, enabling machines to perform complex language tasks with high proficiency. As LLMs become increasingly capable, assessing their true capabilities and limitations becomes crucial. Benchmarks serve as standardized tools to evaluate and compare the performance of LLMs across different tasks. This paper provides a theoretical overview of key benchmarks used to evaluate LLMs, focusing on critical areas such as mathematical and logical reasoning, code generation and programming understanding, multi-task and general language understanding, and commonsense reasoning. By analyzing each benchmark's design, the challenges they present, and their implications, the study aims to shed light on the current limitations of LLMs and suggest directions for future development.

## 2. Background and related work

### 2.1. Evolution of LLMs

LLMs have evolved rapidly, from early models capable of basic language tasks to current architectures that can generate coherent essays and functional code. Scaling model parameters and training data has led to remarkable performance improvements. However, this growth also raises questions about the models' true understanding versus pattern recognition abilities.

## 2.2. The role of benchmarks

Benchmarks are critical for systematically evaluating LLMs' capabilities. They provide means to measure progress, compare models, and identify areas needing improvement. As models become more complex, benchmarks must evolve to present challenges that require genuine understanding and reasoning.

## 2.3. Related work

Previous studies have introduced various benchmarks to evaluate different aspects of LLM performance. Benchmarks like GSM8K [1] and MATH [2] assess mathematical reasoning, HumanEval [3] evaluates code generation capabilities, MMLU [4] examines multi-task understanding, and HellaSwag [5] tests commonsense reasoning. These benchmarks have highlighted limitations in LLMs which will be covered in upcoming sections.

## 3. Benchmark categories and descriptions

### 3.1. Mathematical and logical reasoning

Mathematical and logical reasoning remain significant challenges for LLMs due to the need for multi-step problem-solving and abstract thinking. Two key benchmarks in this category are GSM8K and MATH, which test models' abilities to handle grade-school to competition-level mathematical problems.

#### Grade School Math 8K (GSM8K)

GSM8K, introduced by Cobbe et al. [1], is a dataset of 8,500 high-quality, linguistically diverse grade-school math word problems designed to evaluate LLMs' mathematical reasoning capabilities. Problems range from simple arithmetic to complex multi-step reasoning, typically requiring 2 to 8 steps to solve. Both problems and solutions are presented in natural language, emphasizing the model's ability to provide coherent and understandable explanations.

Despite the dataset's moderate difficulty, even advanced models at the time, such as GPT-3, struggled with GSM8K due to the diversity and complexity of the problems. To address this, Cobbe et al. proposed a "verification" mechanism, training a separate verifier model to assess the correctness of generated solutions. This approach significantly improved performance, highlighting the importance of scalable data approaches and robust reasoning capabilities in LLMs. While newer models have shown improvements, GSM8K continues to serve as a challenging benchmark for assessing mathematical reasoning in LLMs [1].

#### Measuring mathematical reasoning abilities (MATH)

MATH, developed by Hendrycks et al. [2], consists of 12,500 competition-level math problems from middle and high school contests like the AMC (American Mathematics Competitions) and AIME (American Invitational Mathematics Examination). It evaluates advanced mathematical reasoning skills, requiring creative application of concepts and complex multi-step deductions. Each problem includes a detailed step-by-step solution, allowing for an in-depth assessment of the model's reasoning process.

At the time of the original study, fine-tuned models like GPT-3 with 13 billion parameters achieved only 5.6% accuracy on MATH, underscoring the benchmark's difficulty compared to standard NLP (Natural Language Processing) tasks. Providing models with step-by-step solutions during inference often led to decreased performance, indicating challenges in maintaining logical consistency. These results suggested that simply increasing model size and training data was insufficient for significant improvements in advanced mathematical reasoning, emphasizing the need for novel methods to enhance logical reasoning and the ability to generate accurate, coherent solutions [2].

### 3.2. Code generation and programming understanding

Assessing LLMs' ability to generate functional code and understand programming concepts has practical implications for software engineering and automation. The HumanEval benchmark was chosen as an example in this area.

#### HumanEval

HumanEval, presented by Chen et al. [3], is a specialized evaluation set designed to test the functional correctness of Python code generated by LLMs. It consists of 164 unique, hand-written programming problems, each including a function signature, a detailed docstring specifying the task, the function body, and several unit tests. The primary evaluation metric is  $\text{pass}@k$ , measuring the proportion of problems where at least one of the  $k$ -generated samples passes all unit tests, accounting for the stochastic nature of model outputs and the diversity of correct solutions.

At the time of its introduction, results showed that a Codex model with 12 billion parameters solved 28.8% of the problems with one generated sample. Generating more samples significantly increased success rates - up to 77.5% when generating 100 samples and selecting one that passes all tests. Challenges included handling complex instructions, as models struggled with docstrings describing long sequences of operations or mapping operations to variables. Security risks arose when executing generated code, which the study mitigated using a gVisor-based sandbox environment. Additionally, models could generate biased or harmful code, highlighting the need for safeguards against potential misuse [3].

These issues underscore the importance of addressing functional correctness, security, and ethical considerations in code generation tasks. The HumanEval benchmark continues to serve as a valuable tool for identifying persistent challenges and guiding future research in improving LLMs' code generation capabilities.

### 3.3. Multi-task and general language understanding

Evaluating the breadth and depth of LLMs' knowledge across diverse subjects is essential for understanding their general language understanding capabilities. The Massive Multitask Language Understanding (MMLU) benchmark is one of the crucial benchmarks in this assessment.

#### Massive Multitask Language Understanding (MMLU)

MMLU, proposed by Hendrycks et al. [4], comprises over 15,000 multiple-choice questions across 57 subjects, including STEM, humanities, social sciences, and other specialized domains. Questions range from elementary to professional levels, designed to assess models in zero-shot and few-shot settings. Each subject tests aspects such as factual recall, conceptual reasoning, and application of specialized knowledge.

At the time of the study, the largest GPT-3 model achieved an average accuracy of 43.9% in few-shot settings, significantly below the estimated human expert performance of approximately 89.8%. Smaller GPT-3 models performed only marginally better than random guessing, indicating that model size alone was insufficient for high performance without substantial pre-training and fine-tuning. Models like UnifiedQA, with fewer parameters, outperformed GPT-3 in several disciplines, suggesting that pre-training on diverse and well-structured datasets can enhance multi-task language understanding [4].

Observations from MMLU include uneven performance across subjects, with models excelling in areas like U.S. foreign policy but performing poorly in computation-heavy subjects like college chemistry. Calibration issues were noted, where models' confidence levels did not always correlate with accuracy, emphasizing the need for better calibration to prevent over-reliance in critical domains. These findings highlight the importance of diverse knowledge representation and suggest that training paradigms need to evolve beyond simple text prediction [4].

### 3.4. Commonsense reasoning

Assessing LLMs' ability to understand and reason about everyday scenarios requires benchmarks that go beyond surface-level text processing. HellaSwag is often used as a test of commonsense reasoning.

#### **HellaSwag**

HellaSwag, introduced by Zellers et al. [5], is designed to test LLMs' commonsense reasoning abilities using adversarial filtering techniques to create complex and difficult tasks. It consists of approximately 70,000 questions, each presenting a context followed by four possible endings, with incorrect options crafted to be contextually plausible.

At the time of its introduction, despite humans achieving near-perfect accuracy (>95%), models like BERT reached only around 50% accuracy. The models struggled to select the correct ending due to the plausibility of distractors and had difficulty generalizing to zero-shot situations or unseen categories. These results suggested that LLMs at that time lacked robust commonsense reasoning abilities, emphasizing the need for models that could effectively understand and reason about real-world scenarios beyond pattern recognition. The limitations highlighted by HellaSwag pointed to the necessity for further research into training paradigms that foster genuine understanding rather than reliance on superficial cues [5].

## 4. Analysis and discussion

Evaluation across these benchmarks reveals significant limitations in current LLMs. In mathematical reasoning, as assessed by GSM8K and MATH, LLMs struggle with multi-step problem-solving and abstract reasoning; even advanced models at the time like GPT-3 failed to consistently solve moderately difficult problems, indicating deficiencies in logical reasoning that cannot be remedied by merely increasing model size and data.

In code generation, demonstrated by HumanEval, models could produce syntactically correct code but struggle with complex instructions and maintaining functional correctness. Executing generated code poses security risks, necessitating sandbox environments, and models may generate biased or insecure code, raising ethical concerns requiring robust safeguards.

HellaSwag highlights deficiencies in commonsense reasoning; LLMs performed poorly in distinguishing correct endings from plausible distractors, relying on superficial patterns rather than genuine understanding. In multi-task understanding, as evaluated by MMLU, models exhibit uneven performance across subjects, underperforming in computation-intensive or specialized domains. Calibration issues arise, with models overconfident in incorrect answers, posing risks in critical applications.

#### **Common challenges include:**

- Over-reliance on pattern recognition
- Impeding generalization to novel problems requiring authentic reasoning
- Difficulty maintaining logical consistency in multi-step reasoning
- Inaccurate confidence estimations lead to overconfident yet incorrect responses
- Ethical and security concerns due to potential generation of biased, harmful, or insecure outputs

These findings suggest that current training paradigms and benchmarks may not adequately foster deep reasoning and understanding, highlighting the need for innovative approaches prioritizing genuine comprehension over superficial pattern recognition.

## 5. Future directions

Advancing LLM capabilities necessitates enhancing benchmarks and exploring new research methodologies. Benchmarks should become dynamic and adaptive, evolving with model improvements to continuously challenge advanced capabilities, and integrate ethical assessments for bias detection, fairness, and ethical reasoning to ensure responsible AI development. Incorporating multimodal evaluations can assess models in complex, real-world scenarios, promoting more robust understanding.



Research opportunities include improving logical reasoning through novel architectures and training strategies focused on reasoning and problem-solving skills, and emulating human learning processes via interactive learning paradigms to foster deeper understanding and adaptability. Ensuring models align with human values is essential to mitigate risks associated with unintended or harmful outputs.

Collaborative efforts are vital; open collaboration in benchmark development and sharing best practices can accelerate progress, and promoting transparency by sharing datasets, tools, and evaluation frameworks fosters collective advancement and trust in AI technologies. Addressing these areas aims to overcome current limitations, leading to more robust, reliable, and ethically responsible AI systems.

## 6. Conclusions

The theoretical overview of key benchmarks - GSM8K, MATH, HumanEval, MMLU, and HellaSwag - highlights significant limitations in current LLMs. Challenges in mathematical reasoning, code generation, commonsense understanding, and multi-domain knowledge reveal that LLMs often rely on pattern recognition rather than true understanding.

Addressing these limitations requires evolving benchmarks that can adapt to advancing model capabilities and incorporate assessments of ethical considerations. Research focused on improving logical reasoning, exploring interactive learning, and ensuring alignment with human values is essential.

Advancing AI responsibly hinges on collaborative efforts to develop models with genuine comprehension and reasoning abilities. Continuous innovation in evaluation methods and training paradigms will be crucial to overcome current challenges and unlock the full potential of large language models.

### References:

- [1] Cobbe, K., V. Kosaraju, J. Bavarian, M. Bavarian, C. Lu, and J. Zaremba. "Training verifiers to solve math word problems." arXiv preprint arXiv:2110.14168 (2021).
- [2] Hendrycks, D., S. Basart, M. Mazeika, A. Zou, J. Song, and D. Song. "Measuring mathematical reasoning abilities of neural models." arXiv preprint arXiv:2108.09867 (2021).
- [3] Chen, M., J. Tworek, H. Jun, Q. Yuan, H. Ponde, J. Kaplan, and J. Zaremba. "Evaluating large language models trained on code." arXiv preprint arXiv:2107.03374 (2021).
- [4] Hendrycks, D., K. Burns, S. Basart, A. Zou, M. Mazeika, J. Song, and D. Song. "Measuring massive multitask language understanding." arXiv preprint arXiv:2009.03300 (2020).
- [5] Zellers, R., A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. "HellaSwag: Can a machine really finish your sentence?" arXiv preprint arXiv:1905.07830 (2019).

# Retrieval-Augmented Generation (RAG) in Local Large Language Models (LLMs)

Dilan Dobardžić<sup>1</sup>

<sup>1</sup>University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

dilan.dobardzic@tfzr.rs

## Abstract:

Large Language Models (LLMs) have revolutionized various industries by providing advanced natural language processing capabilities, typically delivered through cloud-based deployments. However, many organizations face significant challenges with cloud-hosted models due to concerns over data privacy, sovereignty, and regulatory compliance. This paper explores the implementation of Retrieval-Augmented Generation (RAG) in locally deployed LLMs to address these challenges. RAG enhances LLM functionality by integrating external retrieval mechanisms that access real-time, domain-specific data from internal sources, thereby improving accuracy and relevance while mitigating issues like hallucinations. The paper discusses the key advantages of RAG-enabled local deployments, including enhanced privacy, customizable retrieval for domain-specific needs, and reduced reliance on cloud infrastructure. It also examines the challenges associated with local RAG deployments, such as data leakage risks and computational resource demands, and explores future directions like edge computing integration and advanced security measures. The findings suggest that RAG provides a robust framework for organizations to leverage the benefits of LLMs while maintaining control over sensitive data, ensuring compliance, and improving operational efficiency.

## Keywords:

Large language models (LLMs), retrieval-augmented generation (RAG), local deployments, data privacy, domain-specific retrieval, cloud vs. on-premise

## 1. Introduction

Large Language Models (LLMs) are predominantly hosted in cloud environments where their substantial computational requirements can be managed via scalable infrastructure. Cloud-based deployments offer easy access to powerful models and enable a wide range of applications, from customer service automation to advanced research tools. However, many organizations have raised concerns over privacy, data sovereignty, and regulatory compliance associated with cloud hosting. These concerns have created a need for local LLM deployments. By hosting LLMs on-premise, organizations maintain full control over sensitive data, addressing privacy issues while adhering to strict regulations.

To maximize the effectiveness of LLMs in local settings, Retrieval-Augmented Generation (RAG) has emerged as a valuable technique. RAG enhances the traditional LLM architecture by integrating external retrieval mechanisms that provide access to real-time, domain-specific data from internal sources. This ensures that locally deployed LLMs can generate accurate and relevant responses using up-to-date information without relying solely on pre-trained knowledge. Additionally, RAG helps mitigate issues like hallucinations - where models generate plausible but incorrect information - by grounding responses in factual, retrievable data.

This paper explores how RAG can significantly enhance local LLM deployments by improving data privacy, boosting accuracy, and ensuring regulatory compliance. We will discuss the key advantages of RAG-enabled local deployments, the challenges associated with implementing them, and future directions for this technology.

## 2. Background

### 2.1. Traditional LLM deployments and limitations

Traditional Large Language Models (LLMs) are generally hosted in cloud environments, which provide the necessary computational power for training and inference. Cloud hosting enables seamless scalability and easy access to advanced AI capabilities, making it convenient for a wide range of users. However, cloud-based LLMs face significant challenges, particularly when handling sensitive or proprietary data. Organizations that manage confidential information are often reluctant to transmit sensitive data over public or external networks due to concerns about privacy, data sovereignty, and regulatory compliance.

Relying on cloud-based solutions also introduces potential vulnerabilities, such as data breaches or loss of control over how data is processed and stored. These risks have led organizations to explore local deployments of LLMs as a means of addressing privacy, compliance, and control concerns. By hosting LLMs on-premise, organizations retain full authority over their data, ensuring that sensitive information remains within their secured infrastructure and reducing dependency on third-party cloud providers.

### 2.2. Enhancing local LLMs with RAG

One of the key challenges of deploying LLMs locally is ensuring the model's knowledge remains current and relevant without regular cloud-based updates. Retrieval-Augmented Generation (RAG) addresses this challenge by integrating external retrieval mechanisms that allow the model to access domain-specific or real-time data from local sources. Instead of relying solely on pre-trained knowledge, RAG supplements the model's outputs with fresh, contextually appropriate data retrieved from internal databases or proprietary datasets [1, 2].

RAG operates in two main phases: retrieval and generation [3]. In the retrieval phase, algorithms search for and fetch snippets of information relevant to the user's prompt from a local knowledge base. This knowledge base may include internal documents, policies, or any proprietary data sources that are kept up-to-date. In the generation phase, the LLM utilizes both the retrieved information and its internal knowledge to generate a coherent and contextually appropriate response. This approach allows the model to ground its answers on verifiable and current data, enhancing accuracy and reliability.

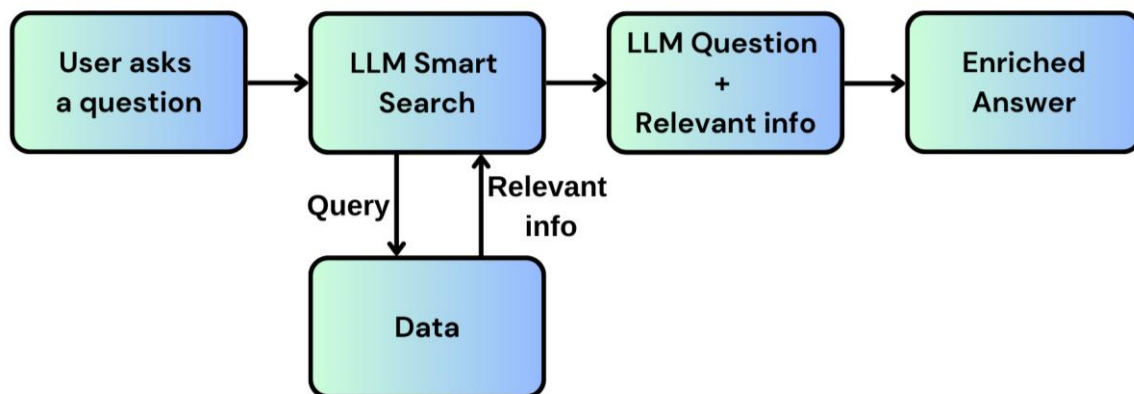


Figure 1: RAG-enhanced LLM workflow

By combining retrieval and generation, RAG offers several benefits for local deployments. It enables LLMs to produce more accurate and context-aware responses by grounding the generated content in verifiable, up-to-date information [3]. This method reduces the likelihood of the model producing hallucinations - a common issue where models generate plausible but incorrect information - by anchoring responses in factual, retrievable data [4]. Additionally, RAG diminishes the need for continuous retraining of the model, as the knowledge base can be updated independently to reflect the most current information.

Implementing RAG in locally deployed LLMs not only enhances the accuracy of responses but also provides transparency into the model's generative process [3]. Users can gain insights into the sources of information the model uses to construct its responses, which is particularly valuable for verifying the accuracy and trustworthiness of the output. This transparency is essential for applications where accountability and traceability of information are critical.

Furthermore, RAG helps mitigate the risk of sensitive data leakage. Since the model relies on external retrieval from secured internal databases, it reduces the chances of inadvertently exposing confidential information embedded in the model's parameters. This is especially important in local deployments where data privacy and regulatory compliance are paramount.

By adopting this "open-book" approach - where the model consults external knowledge bases - organizations can ensure that their LLMs remain current without the need for extensive retraining [3]. This strategy not only conserves computational resources but also allows for rapid updates to the system's knowledge, enabling the LLM to respond accurately to evolving information or policies.

### **3. Benefits of local RAG-enabled deployments**

Implementing Retrieval-Augmented Generation (RAG) in local LLM deployments offers several significant advantages that address common challenges associated with traditional models.

#### **Enhanced privacy and security**

Local RAG deployments keep all sensitive data within the organization's own infrastructure, reducing the risk of data breaches and ensuring compliance with regulations. The retrieval mechanisms access only local databases, ensuring that proprietary or confidential information is never exposed to external servers. Implementing encryption at both the data and vector database levels further safeguards information from unauthorized access [1, 2]. Additionally, limiting reliance on the model's internal parameters reduces the potential for exposing sensitive information, thereby mitigating data leakage risks.

#### **Improved accuracy and relevance**

By grounding responses in up-to-date, domain-specific data, RAG enhances the accuracy and relevance of the outputs generated by LLMs. Accessing real-time information allows the model to provide precise and contextually appropriate responses, which is particularly valuable in contexts where accuracy is critical [3, 5].

#### **Reduction of hallucinations**

Anchoring responses in factual, retrievable data minimizes the risk of the model generating incorrect or misleading information - a phenomenon known as hallucinations [4]. By providing the model with access to verified information, RAG helps ensure that outputs are based on actual data rather than the model's assumptions or inherent biases.

#### **Enhanced transparency and trust**

RAG allows for greater transparency in the model's generative process by providing insights into the sources of information used to construct responses [3]. Users can verify the accuracy of the outputs by referencing the retrieved data, increasing trust in the system. This is particularly important in applications where accountability and traceability are essential.

#### **Customizable domain-specific retrieval**

RAG enables organizations to customize the retrieval component based on their specific needs by fine-tuning it to focus on internal, domain-specific data. Deploying a local LLM that retrieves internal policy documents or proprietary data results in more accurate and tailored responses [3, 5].

#### **Reduced need for retraining**

By accessing up-to-date information through the knowledge base, RAG reduces the necessity for frequent model retraining. Updates can be made directly to the knowledge base, saving time and computational resources [3]. This makes the model more adaptable and efficient in responding to new information without the overhead of retraining.

### **Reduced dependence on cloud infrastructure**

Local RAG deployments significantly reduce reliance on cloud-based infrastructure and constant internet connectivity. In environments with unreliable or restricted internet access, local RAG systems can function seamlessly by leveraging internal resources. This reduced dependency on cloud services is advantageous for organizations operating under strict data control policies or facing connectivity limitations [3].

### **Improved operational efficiency**

By dynamically retrieving up-to-date information without the need for constant retraining, RAG-enabled LLMs improve operational efficiency. Organizations can maintain the relevance of their AI systems while retaining full control over their data, leading to more efficient and effective operations [1].

## **4. Challenges and considerations**

While local RAG deployments offer significant privacy benefits, they also introduce specific security challenges. One of the primary risks is data leakage during retrieval. Since RAG systems retrieve sensitive local data, ensuring that this data is not exposed to unauthorized users is critical. Implementing strict access controls and encryption, particularly for the vector databases used in retrieval, can mitigate this risk [1, 2]. Additionally, organizations need to be cautious about the possibility of adversarial attacks, where malicious actors craft queries to extract sensitive data from the system. Techniques like query validation and content filtering can help prevent these attacks [3, 5].

Another challenge with local RAG deployments is the computational resources required to support both the retrieval and generation components. RAG systems need to maintain large, frequently updated databases to ensure relevant information is always accessible. This can be resource-intensive, particularly for organizations with limited infrastructure. The process of embedding data into vectors and running complex search algorithms like Approximate Nearest Neighbors (ANN) also adds a layer of complexity and requires significant storage and processing power [4, 6].

Keeping the local data source up-to-date is essential for maintaining the relevance of the responses generated by the LLM. However, this comes with the challenge of regularly updating local databases and ensuring the retrieved information is accurate and current. Organizations must implement processes for continuous data validation and ensure that outdated or incorrect information is removed from the system [2, 4].

## **5. Future directions**

As organizations seek to further improve the efficiency of local RAG deployments, the integration of edge computing is becoming a promising avenue. Edge computing enables data processing to occur closer to where the data is generated, reducing latency and improving the responsiveness of RAG systems in local environments. This is particularly relevant in contexts where real-time data retrieval and processing are critical [1].

In addition to edge computing, hybrid models that combine the benefits of both local and cloud-based resources may emerge as viable solutions. These models could enable organizations to offload non-sensitive, high-volume tasks to the cloud while retaining sensitive or critical data locally [3, 6].

Future research in retrieval mechanisms is likely to focus on enhancing the speed and accuracy of the retrieval process, especially for local deployments with large datasets. Techniques such as vector search optimization and the use of more efficient Approximate Nearest Neighbors (ANN) algorithms, like the Hierarchical Navigable Small World (HNSW) algorithm, can help reduce computational overhead and improve search times [2, 4]. Additionally, ongoing advancements in fine-tuning retrieval mechanisms for domain-specific data will further increase the relevance and quality of responses in local RAG deployments [5].

As local RAG deployments become more common, the security of these systems will continue to evolve. Innovations such as homomorphic encryption, which allows computations to be performed on encrypted data, could further protect sensitive information without sacrificing functionality. This will be especially valuable for organizations handling highly sensitive data [5, 6].

Moreover, improvements in query validation and content filtering will help safeguard against adversarial attacks and data leakage, ensuring that RAG systems remain secure while providing high-quality responses [1].

## 6. Conclusions

This paper explored the critical role of Retrieval-Augmented Generation (RAG) in enhancing the performance, privacy, and security of locally deployed Large Language Models (LLMs). By integrating retrieval mechanisms with generative models, RAG addresses some key limitations of traditional LLMs, particularly the problem of hallucination, where models generate factually incorrect or irrelevant information. RAG ensures that responses are grounded in real-time, domain-specific data, which is particularly valuable in contexts that require high accuracy.

A significant benefit of RAG in local deployments is the enhancement of privacy and security. By keeping sensitive data on-premises and away from cloud-based infrastructures, organizations can maintain tighter control over their information, ensuring compliance with relevant regulatory standards. This local control, combined with robust security measures such as encryption and access controls, makes RAG an ideal solution for environments with stringent data protection requirements.

Moreover, RAG's ability to retrieve contextually relevant data from local sources allows organizations to fine-tune their LLMs for specific use cases. This increases the accuracy and relevance of the generated responses, reducing the need for frequent model retraining and improving the overall efficiency of the system.

Additionally, the deployment of RAG in local environments reduces dependency on cloud services, making it particularly beneficial for organizations in remote areas or those with limited internet connectivity.

Looking ahead, advancements in edge computing and hybrid models will likely further enhance the performance of local RAG deployments by reducing latency and improving real-time processing. Ongoing innovations in security, such as homomorphic encryption, will also strengthen data protection without compromising functionality. As retrieval mechanisms become more sophisticated, the efficiency and scalability of RAG systems will continue to improve, making them even more applicable across a wide range of applications.

## References:

- [1] Cloud Security Alliance, "Mitigating security risks in retrieval-augmented generation (RAG) LLM applications," Cloud Security Alliance, 2023. [Online]. Available: <https://cloudsecurityalliance.org/blog/2023/11/22/mitigating-security-risks-in-retrieval-augmented-generation-rag-llm-applications>.
- [2] S. Zeng, J. Zhang, P. He, Y. Xing, Y. Liu, H. Xu, J. Ren, S. Wang, D. Yin, Y. Chang, and J. Tang, "The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG)," arXiv preprint arXiv:2402.16893, 2024. doi:10.48550/arXiv.2402.16893.
- [3] IBM Research, "What is retrieval-augmented generation?," IBM, 2023. [Online]. Available: <https://research.ibm.com/blog/retrieval-augmented-generation-RAG>.
- [4] Rittman Mead, "Retrieval-augmented generation: A primer," Rittman Mead, 2024. [Online]. Available: <https://www.rittmanmead.com/blog/2024/01/retrieval-augmented-generation-a-primer/>.
- [5] P. Lewis, E. Perez, Y. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, L. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," arXiv preprint arXiv:2005.11401, 2020. doi:10.48550/arXiv.2005.11401.
- [6] Elastic, "Retrieval-augmented generation explained," Elastic, 2023. [Online]. Available: <https://www.elastic.co/blog/retrieval-augmented-generation-explained>.

# Optimizing Renal Scintigraphy Interpretation with an Interactive Deep Learning Tool

Olivera Lazić<sup>1</sup>, Marijana Matkovski<sup>1</sup>, Dunja Vrbaški<sup>1</sup>, Ana Jakovljević<sup>2</sup> and Dragan Burić<sup>2</sup>

<sup>1</sup> Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia

<sup>2</sup> Faculty of Medicine, University of Novi Sad, Hajduk Veljkova 3, 21000 Novi Sad, Serbia

*olja.hrnjakovic@uns.ac.rs;*

*marijanamatkovski@uns.ac.rs;*

*dunja.vrbaski@uns.ac.rs;*

*ana.jakovljevic@mf.uns.ac.rs; dragan.buric@mf.uns.ac.rs*

## Abstract:

In nuclear medicine, scintigraphic renal imaging is a diagnostic test used to evaluate the kidney function which utilizes gamma radiation emitted by radioactive tracers. This study introduces a diagnostic interface that integrates human-computer interaction (HCI) principles with adaptive neural networks, improving the interaction between medical professionals and renal scintigraphy images. This application differentiates between healthy and diseased kidney images using the VGG16 neural network, which has been tuned by transfer learning (TL) on a specific dataset of renal scintigraphy images. Its dual-purpose design allows experts to confirm model predictions and feedback while also assisting students in learning diagnostic skills through interactive means. By employing adaptive learning, this feedback dynamically improves the diagnostic accuracy of the model. Early projections show lower mistake rates and higher engagement, underscoring HCI's potential to transform renal scintigraphy interpretation. The purpose of this study is to set a standard for future developments in medical diagnostics that integrate artificial intelligence and HCI, stressing the significance of creating healthcare technologies with the user in mind.

## Keywords:

Adaptive neural networks, HCI, interactive learning, renal scintigraphy, VGG16

## 1. Introduction

Renal scintigraphy, or nuclear renal scan, is a nuclear medicine technique which utilizes different radiopharmaceuticals with high renal clearance. Those radiotracers emit gamma radiation which are captured by gamma cameras. Interpretation of those images plays a crucial role in the evaluation of renal function since it provides critical insights into functional aspects of kidney activity that are not discernible through other imaging techniques [1]. Although there has been a tendency for replacing planar scintigraphy by SPECT (Single Photon Emission Computed Tomography), PET (Positron Emission Tomography) and multimodal imaging such as SPECT/CT, planar scintigraphy is still widely used [2]. Despite its clinical significance, the interpretation of scintigraphy images is a challenging task due to the nature of the data and different protocols, institutional preferences and expert knowledge used for image interpretation [3, 4].

Recent advances in deep learning (DL) has shown great potential for improving diagnostic capabilities of different medical imaging techniques [5]. However, application of advanced computational methods to planar renal scintigraphy is limited. Reasons for this are mainly due to scarcity of large, annotated datasets which could be suitable for neural network training and the nature of data itself. Scintigraphy images have low spatial resolution, lack of depth and very often a surplus of noise which all can be challenging for effective learning. HCI systems leverage the connection between human expertise and machine computational capabilities focusing on interactive solutions that enhance decision-making by combining experts' domain knowledge with the computational power of advanced technologies [6]. Aim of this research is to address challenges in renal scintigraphy images interpretation by developing interactive tool for medical experts and students which will improve diagnostic accuracy. Implementation of adaptive neural network that learn and adjust in real-time, based

on the user feedback, represents a novel integration of HCI principles in interpretation of scintigraphic renal imaging. This paper outlines the development and evaluation of interactive diagnostic tool including data collection, preprocessing and augmentation, use of pre-trained neural networks on small dataset and software design considerations that prioritize usability.

The main contributions of this research are:

- Application of a VGG16 neural network using TL on renal scintigraphy images.
- Implementation of software called NephroLearn, which integrates HCI principles with ANNs to enhance diagnostic accuracy through expert feedback.
- A dual-purpose design in the NephroLearn that offers an interactive teaching tool for student learning while allowing experts to validate model predictions and track student progress.

## 2. Related work

Considering the great influence of technology in today's world in all areas, it is of great importance to monitor how technology affects people through their interaction. One of the most important fields is certainly medicine, and the inclusion of new technologies in medical practice is important for improving diagnostic accuracy, treatment efficiency and overall patient care. The use of artificial intelligence (AI) and virtual reality (VR) can improve clinical outcomes, as well as simplify healthcare processes, making them more efficient and accessible. However, it is very important to monitor HCI in this field, as it ensures that these tools are easy to use, intuitive and effectively meet the needs of both medical professionals and patients [7].

Recent research on the integration of VR into HCI for medical educational tools has shown promising progress in improving medical training and diagnostic accuracy. For example, the use of VR in surgical training that allows trainees to practice without risk leads to better skill acquisition and confidence [8]. Also, a VR-based platform for anatomy education enables medical students to visualize and interact with 3D models of human anatomy, which improves spatial understanding and memory [9]. In addition to the inclusion of VR in medical educational tools, gamification is often used, which involves the application of game design elements in non-game contexts [10]. In this way, user engagement, motivation and learning outcomes can be improved. It has been shown that gamified educational platforms, through the use of challenges, rewards and interactive content, improve the interest of students, and subsequently their performance. Thus, a gamified e-learning tool called TERMINator teaches medical students medical terminology [11]. The tool contains features such as a progress bar, drag and drop exercises, scoring, certificates, and the like. Based on pilot testing with students, it showed positive feedback and improved student results on tests. A study on the inclusion of gamification in a diagnostic tool for radiology training also suggests that gamification can make medical education more interactive and efficient [12]. The study provides users with instant feedback and enables competition with other students, which can lead to diagnostic accuracy and faster learning.

Adaptive learning uses machine learning (ML) algorithms to adjust content and difficulty based on user performance and learning pace. In the context of medical diagnostics, adaptive learning systems can provide personalized feedback and recommendations, thereby improving the accuracy and efficiency of medical professionals. For example, an AI skin cancer detection system that adapts based on diagnostic accuracy and feedback from dermatologists continuously learns and improves its predictions [13]. This approach aligns with best practices for AI and ML in medical imaging, as described in recent recommendations [14]. In this way, better diagnostic results are achieved. Moreover, an adaptive learning framework for radiology education dynamically adjusts the complexity of case studies based on the user's level of expertise, ensuring optimal learning and skill development [15]. In a collaborative review [16], various advanced imaging techniques are presented and compared in terms of their effectiveness in characterizing renal masses. Methods such as multiparametric MRI, contrast-enhanced ultrasound (CEUS) and dual-energy CT show high sensitivity and specificity in distinguishing between benign and malignant kidney lesions. When used together, these imaging modalities provide a comprehensive diagnostic approach with a high degree of accuracy. Among novel imaging tools, 99mTc-sestamibi SPECT/CT shows promise for noninvasive characterization of renal oncocytomas/HOCTs, potentially serving as a comprehensive virtual biopsy in clinical practice. The imaging techniques reviewed in [16] typically use extensive datasets collected from multiple clinical



settings, providing a more comprehensive assessment of renal masses. The findings of this comparative analysis highlight the need to integrate AI-based methods with conventional imaging techniques. The survey on lung CT datasets [17] highlights the importance of large and diverse datasets for robust AI model performance.

These studies highlight the importance of VR, gamification, and adaptive learning in creating intelligent, user-centered medical diagnostic tools that improve both educational and practical outcomes.

### 3. Material and Methods

#### 3.1. Data Collection and Preparation

The dataset that was used in this study is a novel dataset comprising planar renal scintigraphy images. The dataset consists of a total of 24 images, including 10 images with healthy kidneys and 14 images of kidneys exhibiting signs of disease. The images were obtained in cooperation with the Clinical center of Vojvodina in the Clinic for nuclear medicine, respecting ethical guidelines and protocols on patient confidentiality. In order for the images to be standardized as input to the neural network, a detailed preprocessing was done. This included resizing the image to a uniform dimension (256x256 pixels), applying grayscale normalization to accommodate variations in image intensity, and enhancing contrast to improve feature visibility. Such standardization is important because it eliminates discrepancies that may arise due to different settings of the tools on which they were created.

Enhancing the dataset through data augmentation was a crucial step in creating an application that predicts whether an image presents healthy kidneys or not because it uses DL which usually requires large datasets. This was especially important because our dataset was very small. The lack of medical imaging data was a challenge that needed to be addressed in order to prevent overfitting and impair the model's capacity for generalization. By applying numerous alterations to the original photos, data augmentation artificially enhances the amount and diversity of the dataset, hence offering a more robust training environment for the DL model.

We used the ImageDataGenerator class from Keras [18], a powerful feature that facilitates image manipulation during model training. That not only enriches the dataset but also simulates different scenarios under which renal scintigraphy images might be captured, enhancing the model's ability to generalize across unseen images. After several experiments, the following transformations, which included rescaling, rotation (20 degrees), width and height shifts (20%), shear transformations (20%), zoom (20%), and horizontal flipping, showed the best results (Figure 1).

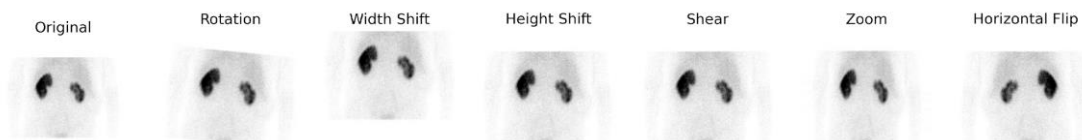


Figure 1: Example of an image from the training set with applied transformations.

#### 3.2. Neural Network Architecture and Training

Unlike traditional ML methods that take advantage of the created features, convolutional neural networks (CNNs) automatically identify the most representative features from the raw input. Some of the main building components that make up a conventional CNN are convolutional, activation, pooling, and fully-connected layers. The core of the NephroLearn is a model made of CNN designed for image classification. The network architecture is based on a pre-trained model, which was fine-tuned for the specific task of classifying renal health from scintigraphic images. We decided to use a pre-trained model because it uses knowledge gained from extensive training on large datasets such as ImageNet and applies it to the specialized task with fewer images.

Training was conducted using a split of 80% of the images for training and 20% for validation. The model was trained with a learning rate of 0.0001 using the Adam optimizer for 30 epochs. We have used the VGG16 model which was modified to exclude its top layers and configured to accept input images of size (256, 256, 3). The VGG16 model was accessed via TensorFlow Hub [19]. We fine-tuned

the model, by unfreezing the last 10 layers. We added a Flatten layer to convert 2D feature maps into a 1D vector, followed by a Dense layer with 512 units and ReLU activation. Batch Normalization was applied to normalize activations, and Dropout with a rate of 0.5 was used for regularization. Early stopping was implemented to overcome overfitting by monitoring the validation loss and halting training if there was no improvement for 20 consecutive epochs. The final output layer was a Dense layer with a single unit and sigmoid activation for binary classification.

With such a small set of data, as we had in the study, it is important to ensure the reliability and robustness of the results. That is why we conducted more training so that we could present the accuracy of the model. Specifically, the VGG16 model was trained 10 times, each with a different random seed to initialize the model weights. After that, we calculated the average accuracy and standard deviation in these series. This approach makes it possible to take into account the variability in the training process and more precisely evaluate the performance of the model. The performance of the model was evaluated using accuracy, recall, precision and F1 score, which are well-known and widely used for classification models.

### 3.3. Interactive Tool

An interactive tool called NephroLearn was developed as an e-learning tool in the form of a website with a dual-purpose design that offers an interactive teaching tool for student learning while allowing experts to validate model predictions and monitor student progress. During the development of NephroLearn, Django 4.2.14 (Django Software Foundation, Lawrence, KS, USA) was used to build the robust web backend, with Python 3.8.10 (Python Software Foundation, Wilmington, DE, USA) as the primary programming language. For neural network we used the Tensorflow framework (version 2.13.1) as backend and Keras (version 2.13.1) as frontend. The frontend was enhanced with Bootstrap (version 4.5.2) for responsive design and jQuery (version 3.5.1) for interactivity. For the database management system, SQLite version 3.31.1 was utilized. In addition, very significant libraries used for developing NephroLearn are NumPy and Pillow.

The NephroLearn interface allows users to load images, view model predictions, and manually adjust or annotate images to provide feedback. Experts are enabled to upload a ZIP archive containing renal scan images. Through the main dashboard they have the opportunity to annotate images integrating further knowledge into the system. The provided information is automatically incorporated into the model through the adaptive learning mechanism. By doing so, the model is dynamically updated and improved. After each input, the tool provides feedback on how well the provided response aligns with the model's response (Figure 2a). Furthermore, experts can follow the learning progress of the model over time. Analytics page shows cumulative advancement over each sequence of images (Figure 3a). Students are provided with an interactive learning framework for practicing interpretation of renal scan images. They are given the opportunity to upload one image at a time, after which they provide their assessment on whether the image shows a healthy kidney function or not. After each entry, the tool provides feedback on matching the given answer with the model's answer (Figure 2b). Clear feedback after each answer helps students improve their skills quickly. After student submission, their daily success rate is automatically updated and available for viewing on a separate page called Results. The daily success rate provides an opportunity to monitor progress and hopefully increases motivation to continue learning. Additionally, experts have the option to review the performance of all students over time (Figure 3b). There is also a separate role for system administrators which can manage roles and have insights into the entire system, hiding unrelatable features from particular users reducing their cognitive load.

NephroLearn was developed with a user-centric design, prioritizing ease of use and intuitive navigation to ensure effective interaction between medical experts and the diagnostic tool. To achieve those goals we followed Schneiderman's eight golden rules of interface design which are: strive for consistency; seek universal usability; offer informative feedback; design dialogs to yield closure; prevent errors; permit easy reversal of actions; keep users in control; reduce short-term memory load [20]. During the development, considerable emphasis was placed on achieving consistency in design. All parts of the application utilize uniform terminology, styles and colors. This principle in user interface design minimizes the need for users to adjust to different parts of the tool while providing a

cohesive user experience. At this point, the application is in its preliminary stage of development and it has a minimum viable, limited, set of features. Nevertheless, design was developed with future use and extension in mind. The user interface has been designed to allow users of all technical skill levels to easily access and use all the functionality intended for them. What should be further addressed is explanations for novices such as tutorials as well as shortcuts or bulk actions for more advanced users aiming to universal usability. Each user action is accompanied by clear and informative feedback. After each user action, notifications are immediately received that clearly indicate the accuracy of their answers and the further action the system takes, depending on the role. In cases where it takes time for the system to process the information, a loading spinner is displayed signaling the user that the action has been accepted and that the content will be displayed soon. The user interface is designed to clearly indicate to users when they have started or finished certain tasks. After each completed activity, the user is offered the option to continue working or move to another activity yielding closure for performed group of actions. The system is designed to minimize the possibility of user error. During data entries, validation mechanisms have been implemented to prevent entering invalid data into the system. In case of an incorrect input attempt, clear messages are displayed to the user, as well as instructions on how to correct them. In this way, the risk of serious errors during use is minimized. As *NephroLearn* is a web tool, the browser's 'undo' and 'redo' system allows for effortless reversion to previous steps with one consideration, model adaptive learning. There is no option for a model to forcefully redo learning. This is a somehow challenging task, not only in implementation but also considering its validity and necessity. At this point, if it is applicable, the user is informed that by using the browser “back” feature will not cancel what model has learned by last action. The system is designed to allow users to feel in control of their work. It reacts to user actions in a predictable and familiar way with no space for surprises or challenges in use. This is mainly due to a fixed set of features with uniform design and clear feedback. As the tool progresses through future versions and new features implementation this aspect of good user design should especially be considered and monitored. Due to software nature, goals and tasks, *NephroLearn* does not have the issue with unnecessary loading of the user’s short-term memory. There are no complicated forms or connected displays, design is simple and transparent, thereby making the system easy to use with not much information which the user should remember or reuse. Addressing all the rules for good user interface design we have set strong grounds for *NephroLearn* to be used by students and experts and to evolve efficiently in subsequent versions.

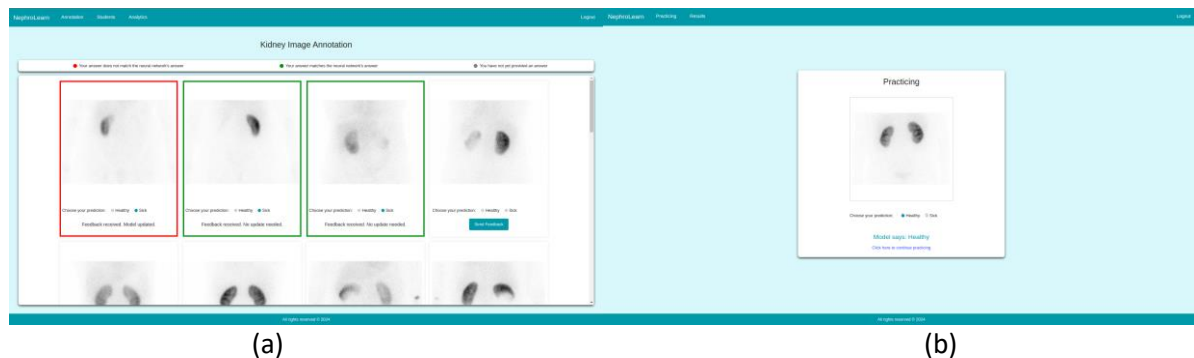


Figure 2. Expert annotation and feedback and student practice and feedback in NephroLearn

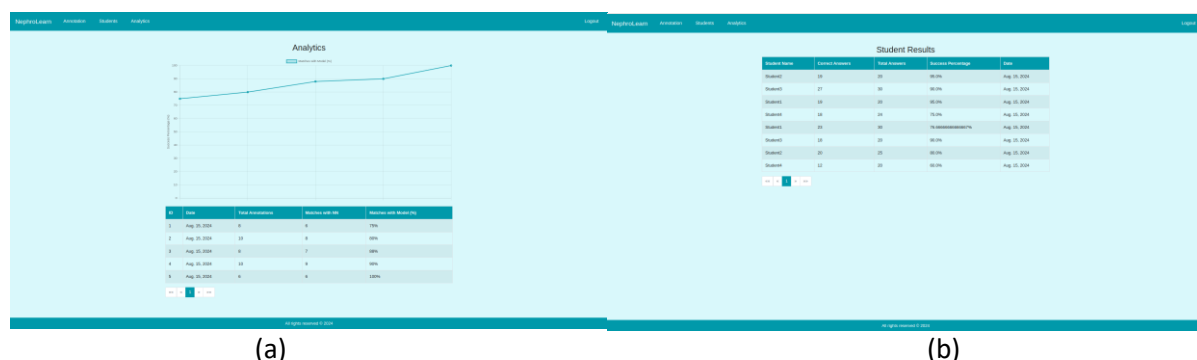


Figure 3. Analytics dashboard and experts access to student performance data in NephroLearn.

## 4. Results and Discussion

### 4.1. Results

The applied CNN model achieved an average accuracy of 0.78 with a standard deviation of 0.23 over 10 training sessions (Figure 4). This cross-training approach helped mitigate the effects of random initialization and data mixing. In this way, a more reliable measure of model performance was obtained. The model with average accuracy of 0.8 was evaluated using recall, precision and F1 score, achieving solid results, with recall of 1.0, precision of 0.75 and F1 Score of 0.86.



Figure 4: The accuracy of VGG16 model for each run.

### 4.2. CNN potential in interpretation of scintigraphy images

Our study uses the neural network model VGG16, a type of CNN, to classify renal scintigraphy images into healthy and diseased categories. Despite the limited data set of 24 images, the model achieved an accuracy of 0.8. This performance is promising, given the data size limitations and inherent complexity of image modality used. The performance indicators and diagnostic accuracy of 0.8 obtained by the VGG16 model indicate significant potential for AI-based methods in renal health detection. Although this accuracy is somewhat lower compared to the studies on similar medical images, it is important to consider the context of the data set and differences in methodology. There is no other research that focuses on renal scintigraphy images which makes this work unique.

One of the critical challenges in our study is the limited dataset size of 24 images, with only 10 images showing healthy kidneys and 14 showing signs of disease. The small sample size may affect the generalizability and robustness of the model. Even so, the neural network's ability to learn and identify complex patterns in scintigraphy images suggests that with larger datasets and further refinement, AI models could achieve comparable diagnostic accuracy to traditional imaging methods. Future research should focus primarily on increasing the size and diversity of the dataset and incorporating multimodal imaging data to improve the diagnostic accuracy and generativity of the model. In this way, we could develop a more robust and comprehensive diagnostic tool for characterizing the renal mass.

### 4.3. Interactive tool for fostering adoption of AI models

Interactive tool, as an integral part of the research effort in this study, was developed to effectively engage medical professionals in using AI powered tool. Such approach not only facilitate more accurate interpretation by both model and humans but also foster student learning and empower professionals to participate in further development. Interactivity of the tool, with the real-time updating and feedback integration, enables professionals to have effortless hands-on experience with AI model without the burden of unnecessary cognitive load due to unfamiliar technology.

The tool was designed following the rules of good user interface design. Although every rule was addressed and taken into consideration during the design there is a place for improvement. Firstly, future versions should consider errors and adversities that can arise from internal implementation and mechanisms, not only users. For example, we must consider on what conditions model training could be unresponsive or can it be “hacked” in order to give erroneous cumulative results for students. Also, considering that the main task of the tool is the interpretation of the renal scan images, it would be preferred if the tool has some information and explanations about model interpretations. That

information will raise the trust in the model and give opportunity for experts to comment, discuss and influence prospective improvements. The AI model should act as black box, integrated in the tool, but some reasoning should be provided through implementation of explainable AI mechanisms. Development of HCI tools, for a specific group of users, could be challenging and should be accessed with caution, considering all aspects of user interface design in every step of software development. In the era where many applications are being AI powered, new challenges arise and the concept of “Human-centered AI” (HCAI) is introduced [21]. Xu et al. provides a comprehensive overview of the challenges and the opportunities for HCI professionals which are specifically related to AI driven applications since those systems exhibit unique characteristics [22]. In order to design a future mature, usable and reliable system we have to address all those issues including explainability, ethics questions and full and seamless integration and collaboration of human and machine intelligence.

Therefore, future extensions will be aimed at both model improvements and tool enhancements, following not only standard HCI best practices but also considering HCAI approaches. With those advancements, controlled and systematic testing of the user interface by students and experts should be continuously performed in order to measure efficiency of different interface design tasks and enable stronger human controlled AI.

## 5. Conclusion

Our approach using the VGG16 neural network and TL shows promising accuracy in the classification of scintigraphy renal images despite the limitation of a small dataset. However, to achieve further progress, it is necessary to expand the dataset and integrate these methods with existing imaging methods. In this way, a reference value for the performance of advanced imaging techniques would be obtained, which could serve as a valuable reference point for the continuous development of diagnostic tools based on AI. Implementation of interactive NephroLearn software tool enables the model to be practically usable by both medical experts and students, bridging the gap between scientific capabilities of AI models and feasible use, enhancing trust and adoption of AI in clinical settings and learning environments.

### **Ethical Approval:**

This research was approved by Ethical Committee of the Clinical Centre of Vojvodina, Republic of Serbia (No. 00-299). All procedures performed in the study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the Declaration of Helsinki.

### **Funding:**

This research was supported by the Ministry of Science, Technological Development, and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad, by the “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad” project (No.01-3394/1).

### **References:**

- [1] M. D. Blafox et al., "The SNMMI and EANM practice guideline for renal scintigraphy in adults," *Eur J Nucl Med Mol Imaging*, 45(12), 2218–2228, 2018. <https://doi.org/10.1007/s00259-018-4129-6>.
- [2] M. Gotthardt, C. P. Bleeker-Rovers, O. C. Boerman, W. J. G. Oyen, "Imaging of Inflammation by PET," *J Nucl Med*, 51(12), 1937–1949, 2010. <https://doi.org/10.2967/jnumed.110.076232>.
- [3] A. T. Taylor, "Radionuclides in Nephrourology, Part 1: Radiopharmaceuticals," *J Nucl Med*, 55(4), 608–615, 2014. <https://doi.org/10.2967/jnumed.113.133447>.
- [4] A. T. Taylor, "Radionuclides in Nephrourology, Part 2: Diagnostic Applications," *J Nucl Med*, 55(5), 786–798, 2014. <https://doi.org/10.2967/jnumed.113.133454>.
- [5] M. Zhang et al., "Imaging-based deep learning in kidney diseases," *Insights Imaging*, 15(1), 50, 2024. <https://doi.org/10.1186/s13244-024-01636-5>.

- [6] L. Rundo et al., "Advances of HCI in decision-making for clinical workflows," *J Biomed Inform*, 108, 103479, 2020. <https://doi.org/10.1016/j.jbi.2020.103479>.
- [7] R. Mishra, R. Satpathy, B. Pati, "Interpretable AI in Medical Imaging," *AIS*, 6(1), 96–111, 2024. <https://doi.org/10.33969/AIS.2024060107>.
- [8] J. Pottle, "Virtual reality and medical education," *Future Healthc J*, 6(3), 181–185, 2019. <https://doi.org/10.7861/fhj.2019-0036>.
- [9] R. Barmaki et al., "AR in anatomical education," *Anat Sci Educ*, 12, 2019. <https://doi.org/10.1002/ase.1858>.
- [10] H. Kandemir, H. Kose, "Adaptive HCI games to evaluate attention," *Robotica*, 40(1), 56–76, 2022. <https://doi.org/10.1017/S0263574721000370>.
- [11] A.-H. Seidlein et al., "Gamified E-learning in medical terminology," *BMC Med Educ*, 20(1), 284, 2020. <https://doi.org/10.1186/s12909-020-02204-3>.
- [12] M. F. Ali et al., "Gamification in radiology training," *BMC Res Notes*, 14(1), 341, 2021. <https://doi.org/10.1186/s13104-021-05761-y>.
- [13] A. Esteva et al., "Dermatologist-level skin cancer classification with DNNs," *Nature*, 542(7639), 115–118, 2017. <https://doi.org/10.1038/nature21056>.
- [14] L. Hadjiiski et al., "AAPM Task Group Report 273," *Med Phys*, 50(2), e1–e24, 2023. <https://doi.org/10.1002/mp.16188>.
- [15] C. Liu et al., "Paediatric radiology escape room," *Insights Imaging*, 11(1), 50, 2020. <https://doi.org/10.1186/s13244-020-00856-9>.
- [16] E. Roussel et al., "Novel imaging for renal mass characterization," *Eur Urol*, 81(5), 476–488, 2022. <https://doi.org/10.1016/j.eururo.2022.01.040>.
- [17] R. V. Adiraju, S. Elias, "Survey on lung CT datasets," *Res Biomed Eng*, 37(2), 403–418, 2021. <https://doi.org/10.1007/s42600-021-00138-3>.
- [18] F. Chollet, Keras, GitHub, 2015. <https://github.com/fchollet/keras>.
- [19] TensorFlow Hub, TensorFlow, 2024. <https://www.tensorflow.org/hub>.
- [20] B. Shneiderman et al., *Designing the User Interface*, 6th ed., Pearson, 2016. <http://www.cs.umd.edu/hcil/DTUI6>.
- [21] B. Shneiderman, "Human-Centered AI," *Int J Hum Comput Interact*, 36(6), 495–504, 2020. <https://doi.org/10.1080/10447318.2020.1741118>.
- [22] W. Xu et al., "Human interaction with AI systems," *Int J Hum Comput Interact*, 39(3), 494–518, 2023. <https://doi.org/10.1080/10447318.2022.2041900>.

# Techniques of Software Integration Testing

Zoltan Kazi<sup>1</sup>, Ljubica Kazi<sup>1</sup>, Madhusudan Bhatt<sup>2</sup>, Narendra Chotaliya<sup>3</sup>

<sup>1</sup> University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Republic of Serbia

<sup>2</sup> University of Mumbai, K.C. College, India (retired)

<sup>3</sup> Saurashtra University, Government Science College, Jasdan, Rajkot, India

*zoltan.kazi@tfzr.rs; ljubica.kazi@gmail.com; mmbhatt@gmail.com; narendra\_chotaliya@yahoo.com*

## Abstract:

Software quality control is a very important set of activities in software development process. It can save money, increase security of applications and user data, increase customer satisfaction and achieve overall adequate product quality. The need for quality software is enormous. This paper describes integration testing strategy and shows the most important techniques of software integration testing on an abstract example. The importance and position of integration testing in the entire process of software development and testing is highlighted.

## Keywords:

Software, testing, integration, techniques

## 1. Introduction

In the recent years information systems have become increasingly complex - networked software systems that are increasingly used within many human activities. The complexity always induces risks. On the other side, there is a high competitiveness in software production. As a consequence, developed software must retain maintainability and testability, while at the same time, very high reliability must be ensured. Otherwise, the consequences can be catastrophic. [2]

The need for quality software is challenging today. It is estimated that in practice from at least 30% to even 50% of the total effort and costs in software development goes to testing and quality control [4]. Malfunctioning software causes user dissatisfaction and financial, material, loss, or even human casualties. Therefore, software quality control is a very important set of activities in its development. Software testing needs to be done to save money, increase security of applications and user data, increase customer satisfaction and achieve adequate product quality. Software testing is an integral part of Software Development Life Cycle (SDLC) process [3], [5]. In SDLC, the software is finished after it has passed testing. The earlier errors are detected, it is cheaper to fix it.

The main purpose of testing is to provide confidence that the software is working well. Good testing provides measures for the following relevant factors: correctness, efficiency, flexibility, reliability, testability, reusability, usability, documentation, maintainability, and integrity structure [7]. One of the methods that is insufficiently applied in software quality assurance is integration testing.

The rest of the paper is organized as follows: section two presents integration testing strategy; section three integration testing techniques, section four integration testing techniques on abstract example; and final section brings the conclusion.

## 2. Integration testing strategies

Testing strategies are related to levels of software system complexity and are categorized as [5]:

- Unit testing
- Integration testing
- System testing

The classification of software testing into levels can be seen on Figure 1. Based on the client's needs, the specification of the user's requirements is formed. Besides that the planning and design of

the software, its' elements and architecture are carried out. Based on the design, programs are written and the first tested parts are code with unit tests. After that, the parts of the program must be connected into larger units, modules and integration testing is done based on the previous design, models and different specifications. This is followed by system testing of the entire software to see if it complies with the specification, and at the end, acceptance tests are performed to determine whether the software meets its purpose, the basic task that the user needs. [8]

After determining the correctness of certain components, i.e. of unit testing of the basic parts of the code, they are integrated into larger units of the software system that should function according to the set requirements. Integration testing checks unit modules that are combined into larger units and how they work together. The biggest problem when integrating components is their connection through software interfaces. Interfaces between components can lead to the following problems:

- Data can be lost when passing through the interface between components,
- One component can have a negative impact on another,
- Combinations of program sub-functions may have an adverse effect on the main function,
- Inaccuracies in programming, acceptable at the level of an individual component, after connection give unacceptable values (eg variable and more complex data structures).

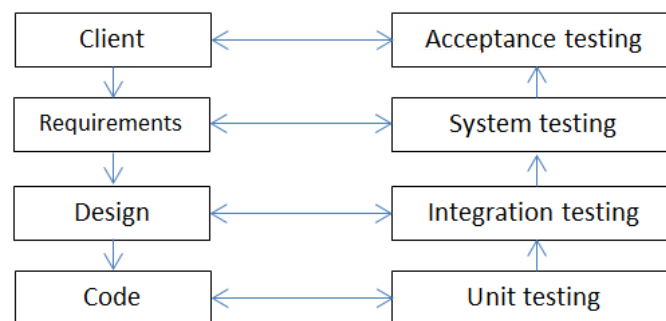


Figure 1: Software testing levels [6]

The integration testing strategy defines how and in what way the components are combined to test the system, which affects the integration time schedule, the coding order, and the cost and comprehensiveness of testing. The system is viewed as a hierarchy of components, in which each component belongs to a design layer. The goal of integration testing is detailed testing of the integration and interfaces between components, as well as the connection with the operating system, file system, hardware and software interfaces, etc.

### 3. Integration testing techniques

According to [1] the main traditional integration testing strategies are: “top-down” integration, “bottom-up” integration, “big-Bang” integration, threads integration, and critical modules integration.

According to [1] main integration testing techniques are:

- Integration from top to bottom ("Top-Down"),
- Integration from the bottom to the top ("Bottom-Up"),
- Integration according to the "Big Bang" principle,
- Sandwich integration,
- Threads integration.

In the following, we will address and explain some of the mentioned techniques of integration testing.

Top-Down integration is a favorite method of many designers and testers. At the top level, there is usually one component that is tested in isolation. After that, components that are used by previously tested component are combined and tested as larger units. This process is repeated until all components are included. A component being tested may call another that is not yet tested, so a dummy routine (“stub”) is created, a special-purpose program that simulates the activity of the component that is missing because it is still under development. Stab pretends to be a non-existent



module, responds to the call, accepts input data, and returns output data so that the testing process can continue. A stub is a component that simulates the operation of a lower-level module, and a driver is a component that simulates the operation of a higher-level module. [1] The disadvantage of this type of testing is the potential need for a large number of fake routines. This can happen if the lowest level of the system contains a large number of general purpose routines. One way to avoid this problem is to change the strategy. Instead of including an entire tier at a time, in a modified top-down approach, components of each tier are unit tested before merging. Unit testing the components of each level causes another difficulty. Each component requires dummy routines and handlers, resulting in more coding and therefore more potential problems.

When using Bottom-up integration testing method, each component at the lowest level of the system hierarchy is the first unit that is tested. Then there are the component tests that call the previously tested components. This procedure is repeated until all components have been tested. The Bottom-Up method is useful when most of the lowest-level components represent general-purpose utility routines that are called from other components, as in object-oriented design, or when the system integrates a large number of independent reusable components. [1] The disadvantage of this type of testing is that it is difficult to test the system from the bottom up if there is a time dependency in a large part of the processing, in the system. Often this technique is best suited for object-oriented programs. Objects are combined one by one with objects or collections of previously tested objects.

"Big Bang" integration or "joining components at once" is also used within integration testing. After the unit testing of all components, the final integration of the system follows and checking whether everything works as specified and designed. In 1979, the author, Myers, called it "Big Bang" integration. Many testers use this method on smaller systems, while its' application on large systems is impractical with potential problems. The first problem is that in order to test individual components, it is necessary to write both stubs and drivers. Another problem is that, since all components are integrated at once, it is difficult to find the cause of irregularities, and errors in interfaces are difficult to separate from other types of errors[1].

Sandwich integration was created as a combination of techniques from top to bottom and from bottom to top. The author of sandwich integration is Myers, in 1979. According to him, the system is viewed in three layers, like a sandwich. The target layer is in the middle, and there is one layer each on the bottom and top sides. A top-down approach is applied to the top layer, and a bottom-up approach to the bottom layer. Testing approaches a target layer that is selected based on system characteristics and the structural hierarchy of components. Stubs and drivers are required for a small number of modules/components. [1] This type of integration allows testing to be done as early as possible during software implementation. Sandwich integration has an advantage over top-down and bottom-up approaches by testing control components and utilities at the outset. There is no comprehensive testing of individual components before integration. One variation, modified sandwich testing, involves testing higher-level components before merging them with others. This is a hybrid form of integration, a combination of top-down and bottom-up approaches [7]. The advantages of this approach are: Suitable for large projects, Fewer stubs and drivers are needed, Executes faster than other approaches. Deficiency is related to the fact that defect isolation is more difficult.

Threads integration strategies can be applied to object-oriented systems. Object-oriented systems are sets of cooperating entities that exchange messages, and threads can be identified with sequences of subsequent message invocations. Therefore, a thread is a scenario of normal usage of an object-oriented system. Testing a thread implies testing interactions between classes according to a specific sequence of method invocations. [1]

## 4. Example

Example of program modules, labeled: A, B, C, D, B1, B2, D1, D2, D3, B21, B22 that will illustrate integration testing techniques (Figure 2). Software basic component A consists of three main components: B, C, D. Component B consists of B1 and B2, while B2 consists of B11 and B12. Finally, module D has sub-modules: D1, D2 and D3.

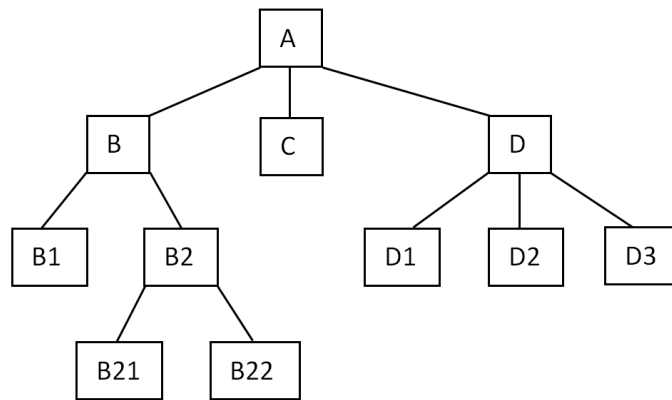


Figure 2: Program modules example

Integration from top to bottom ("Top-Down") - Figure shows how to test the system using the top-down method. Component A, at the highest level, is the only one that is tested in isolation, while B, C, and D require dummy routines. When component A is tested, it is combined with the next level so that B, C and D are tested together. At this stage of testing, dummy routines for B1, B2, D1, D2, D3 will be needed. Finally, the entire system is tested.

Test level 1: testing A (1. on Figure 3)

Test level 2: testing  $A > B > C > D$  (1.+2. on Figure 3)

Test level 3: testing  $A > B > B1 > B2 > C > D > D1 > D2 > D3$  (1.+2.+3. on Figure 3)

Test level 4: testing  $A > B > B1 > B2 > B21 > B22 > C > D > D1 > D2 > D3$  (1.+2.+3.+4. on Figure 3)

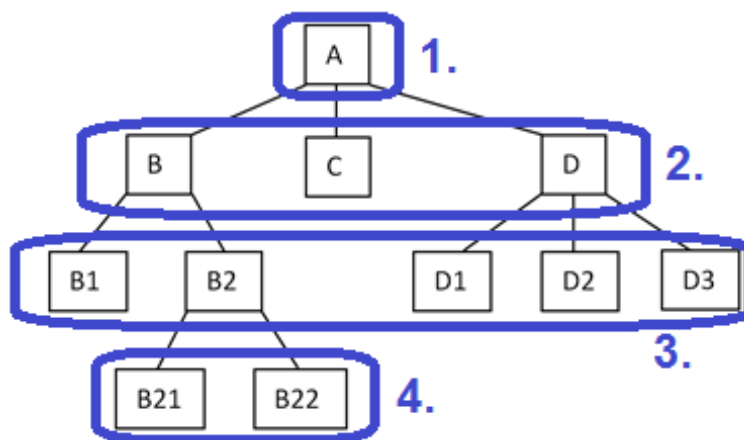


Figure 3: Integration from top to bottom

A system can be tested using a modified approach by first testing A, then B, C, and D, and only then integrating those four components for level one and two level testing. Then B1, B2, D1, D2, D3 are tested individually. Finally, the entire system is brought together for testing.

Integration from the bottom to the top ("Bottom-Up"):

Test level 1: testing  $B21 > B22$  (1. on Figure 4)

Test level 2: testing  $B1 > B2 > B21 > B22 > D1 > D2 > D3$  (1.+2. on Figure 4)

Test level 3: testing  $B > B1 > B2 > B21 > B22 > C > D > D1 > D2 > D3$  (1.+2.+3. on Figure 4)

Test level 4: testing  $A > B > B1 > B2 > B21 > B22 > C > D > D1 > D2 > D3$  (1.+2.+3.+4. on Figure 4)

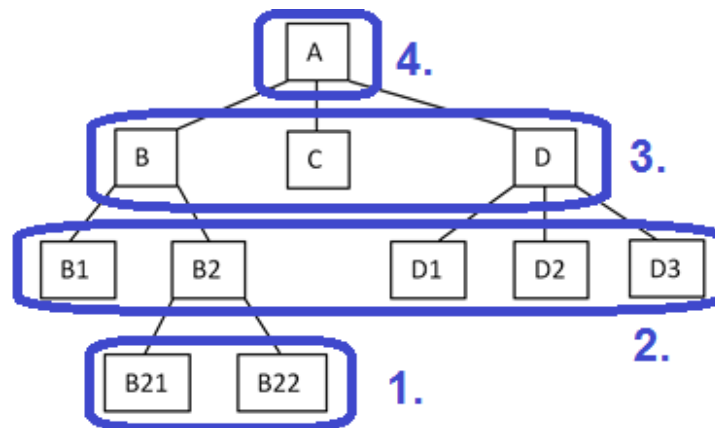


Figure 4: Integration from the bottom to the top

Integration according to the "Big Bang" principle:

Test level 1: testing  $B > B1 > B2 > B21 > B22 > C > D > D1 > D2 > D3$  (1. on Figure 5)

Test level 2: testing  $A > B > B1 > B2 > B21 > B22 > C > D > D1 > D2 > D3$  (1.+2. on Figure 5)

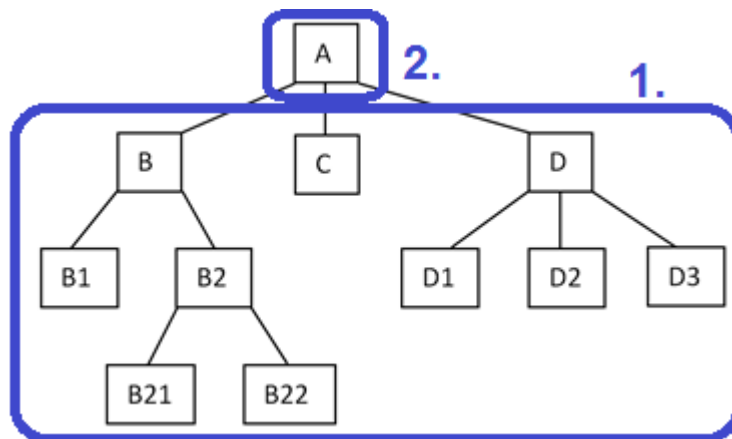


Figure 5: "Big Bang" integration

Sandwich integration - (like eg. "Test A" in the picture below, a stub is required for A). The following figure shows the sequence of tests for sandwich integration, where the target layer is in the middle, i.e. components: B, C and D.

Test level 1: testing  $B1 > B2 > D1 > D2 > D3$  (1. on Figure 6)

Test level 2: testing  $B > B1 > B2 > C > D > D1 > D2 > D3$  (1.+2. on Figure 6)

Test level 3: testing  $B > B1 > B2 > B21 > B22 > C > D > D1 > D2 > D3$  (1.+2.+3. on Figure 6)

Test level 4: testing  $A > B > B1 > B2 > B21 > B22 > C > D > D1 > D2 > D3$  (1.+2.+3.+4. on Figure 6)

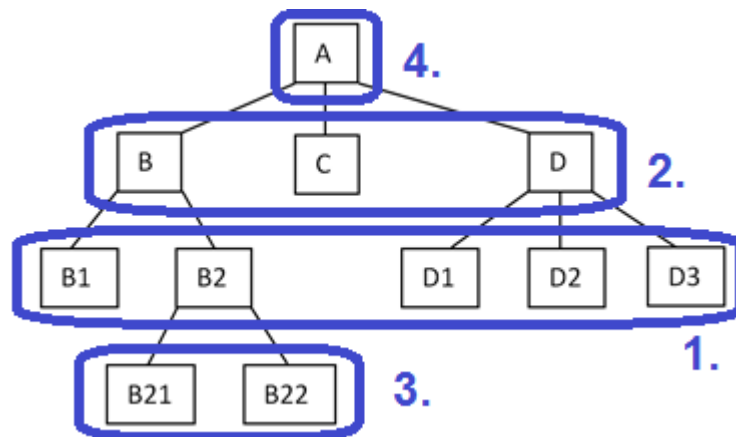


Figure 6: Sandwich integration

## 5. Conclusions

Software quality control is a very important set of activities in its development. Integration testing is not given enough attention and in practice, it is often not done satisfactorily, which can lead to certain problems. The mentioned techniques of integration testing are: integration from top to bottom (top-down), integration from bottom to top (bottom-up), "Big Bang" integration, Sandwich integration, and Threads integration.

We can conclude that the top-down integration is a favorite method of many designers and testers; "Big-Bang" method is suitable on smaller systems; Sandwich method is suitable for large projects; Threads integration strategies can be applied to object-oriented systems; and the bottom-up (bottom-up) method is useful for lowest-level components and general-purpose utility routines that are called from other components, also in object-oriented design, or when the system integrates a large number of independent reusable components.

From the above we conclude that it is very important to know well the techniques of integration testing in order to properly apply them in certain software, which, of course, are not all of the same character, purpose, complexity, etc.

Future direction in this field could be in development of tools that could combine different integration testing strategies and methods, as well as to combine integration testing with other types of testing methods. Results of these tests could be stored and used by developers not only to fix code, but to learn what is not to be done in future developments. In research, it would be beneficial to conduct systematic review of papers related to testing methods, to organize taking questionnaire among testers regarding their work with different methods and tools, or to develop methods and tools that are based on accumulated testing results, that could be used in predicting possible future bugs, as a guidance to novice software developers.

### References:

- [1] A. Orso, Integration Testing of Object-Oriented Software, Politecnico Di Milano, PhD Thesis, 2021.
- [2] M. Rehman, F. Jabeen, A. Bertolino, A. Polini, Testing Software Components for Integration: A Survey of Issues and Techniques, *Softw. Test., Verif. Reliab.* 17 (2007): 95-133. 10.1002/stvr.357.
- [3] J. E. T. Akinsola, M. A. Adeagbo, S. O. Abdul-Yakeen, F. O. Onipede, and A. A. Yusuf, Qualitative Comparative Analysis of Software Integration Testing Techniques, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 7 No. 2, (2022) pp. 67 – 82.
- [4] D. Drašković, D. Bojić, Software Testing, Akademska misao, Belgrade, 2019. (Serbian)
- [5] A. M. Umar. "A Study of Software Testing: Categories, Levels, Techniques, and Types", *TechRxiv.*, June 30, (2020) DOI: 10.36227/techrxiv.12578714.v2
- [6] M. Živković: "Software Testing", Singidunum University, Belgrade, 2018. (on Serbian)

- [7] P. Jiantao: "Software Testing", Carnegie Mellon University, 18-849b Dependable Embedded Systems, Spring 1999.
- [8] Z. Kazi, M. Kazi, Software Testing Strategies, Approaches, Methods and Techniques - Overview, 13th International conference on Applied Internet and Information Technologies AIIT2023, October 13, 2023, Bitola, Republic of North Macedonia, pp. 321-328.

# AI at the Edge: Trends and Innovations in Tiny Machine Learning Models for IoT and Embedded Systems in Synergy with Neuton.AI

Aneta Trajkovska<sup>1</sup> and Aleksandar Markoski<sup>1</sup>

<sup>1</sup>University “St. Kliment Ohridski”, Faculty of Information and Communication Technologies, Bitola, Republic of Macedonia

aneta.trajkovska@uklo.edu.mk, aleksandar.markoski@uklo.edu.mk

## Abstract:

The trajectory of technological evolution is increasingly oriented towards the development of intelligent solutions that enhance both the efficiency and functionality of everyday life. As technological advancements accelerate, we are witnessing a paradigm shift in the execution of technical processes, aimed at simplifying device interactions while simultaneously enhancing control and automation. The rise of AI at the edge is revolutionizing the way we approach machine learning in the context of IoT and embedded systems. Edge AI, which brings the power of machine learning to edge devices, allows for real-time data processing and decision-making, enabling devices to operate independently of cloud-based systems. This innovation is crucial for applications requiring low-latency responses, such as autonomous vehicles, smart cities, and industrial automation. The convergence of AI, IoT, and edge computing is thus driving significant innovation in embedded systems, with trends indicating a growing emphasis on lightweight machine learning models, energy-efficient algorithms, and scalable architectures. In this paper, we will conduct an in-depth exploration of the utilization of TinyML systems, focusing particularly on practical case studies and best practices associated with neuton.ai. By examining practical use cases of neuton.ai, we will highlight its contributions to advancing the field, including innovations in model optimization, scalability, and real-world deployment strategies.

## Keywords:

Internet of Things, AI, Edge intelligence, neuton.ai, tiny machine learning,

## 1. Introduction

The evolution of the Internet of Things (IoT) can be framed as an integral part of the successive industrial revolutions that have transformed societies and economies. Each phase of industrial development has seen the introduction of new technologies that enhance productivity, efficiency, and connectivity [1],[2]. IoT, as we know it today, can be understood as a product of the Fourth Industrial Revolution, but its foundational concepts trace back to earlier transformations in industrial history Figure 1. Each revolution brought about a new paradigm of connectivity and automation, and IoT represents the latest stage in this trajectory. By integrating data, devices, and systems, IoT is not only enhancing industrial processes but also transforming how humans interact with the world around them. The convergence of IoT and AI is driving a new wave of industrial and societal transformations, enhancing system capabilities and human-machine collaboration. [3].

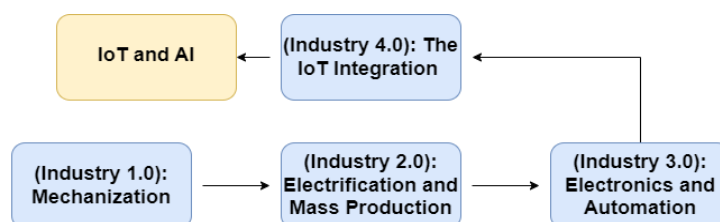


Figure 1: The four industrial revolutions and current progress

The integration of machine learning into resource-constrained devices has led to the emergence of TinyML, a field dedicated to enabling intelligent data processing and decision-making at the edge. As the Internet of Things (IoT) continues to expand, the demand for efficient, low-latency, and energy-conscious AI solutions has become critical. TinyML addresses these demands by allowing machine learning models to operate on small, embedded devices with minimal computational power, reducing the need for continuous cloud connectivity and optimizing real-time performance [4].

One of the leading innovations in this domain is neutron.ai, which provides tools and frameworks for developing highly efficient TinyML models without the complexity of traditional machine learning pipelines [5],[6]. By automating model generation and optimizing performance for edge devices, neutron.ai has demonstrated practical applications across various industries, including healthcare, agriculture, and industrial automation. These advancements allow for real-time data analysis and decision-making directly on the device, minimizing latency, bandwidth usage, and energy consumption.

This paper explores the practical utilization of TinyML systems, with a specific focus on neutron.ai's capabilities [7]. It aims to provide a detailed analysis of current trends, challenges, and best practices for deploying TinyML in embedded systems, addressing the implications for both technological innovation and real-world applications.

## 2. Technological Trends and Innovations

Technological trends and innovations in TinyML are driving significant advancements in the fields of IoT and edge computing. One major trend is the development of increasingly efficient hardware, such as low-power microcontrollers and specialized AI accelerators, which enable real-time data processing directly on edge devices. Model optimization techniques, including quantization and pruning, have also emerged as key innovations, allowing machine learning models to operate with minimal computational and energy requirements. Platforms like neutron.ai are advancing the automation of TinyML model creation, enabling the deployment of highly optimized models on resource-constrained devices without requiring deep expertise in machine learning Figure.2.

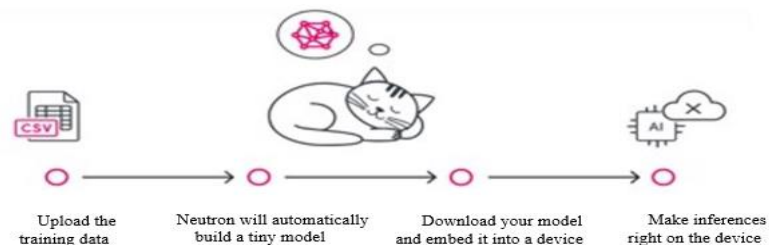


Figure 2: Neutron.AI high architecture process overview [8]

Moreover, innovations in lightweight neural architectures and energy-efficient algorithms are addressing the challenges of deploying AI on embedded systems [9]. These trends are expanding the applicability of TinyML across diverse sectors, from healthcare and smart cities to industrial automation, by enabling real-time, decentralized decision-making. The convergence of these innovations is shaping the future of AI at the edge, allowing for more scalable, secure, and intelligent IoT ecosystems.

### 2.1. IoT and AI integration for embedded systems

A key development is the advancement of energy-efficient hardware, such as microcontrollers and edge processors, designed specifically to handle machine learning tasks directly on devices with limited computational resources. This evolution is critical for real-time, decentralized data processing, reducing reliance on cloud-based infrastructures and lowering latency. In parallel, model optimization techniques like quantization, pruning and neural architecture search have enabled the deployment of complex AI algorithms on constrained devices. Platforms such as neutron.ai are playing a pivotal role in automating the creation and deployment of highly efficient TinyML models, allowing embedded systems to execute advanced AI tasks without requiring extensive expertise or large computational overhead [10],[11].

The integration of AI and IoT in embedded systems has led to innovations in autonomous decision-making, particularly in industries such as healthcare, agriculture, and industrial automation. The ability to process data locally enables enhanced privacy, reduced bandwidth usage, and faster response times, making these systems highly adaptive for real-world applications Figure 3. As these trends continue, the convergence of IoT, AI, and TinyML is shaping the future of intelligent, low-power embedded systems, offering scalable and efficient solutions for next-generation smart environments [12].

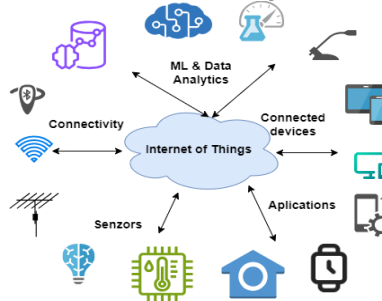


Figure 3: IoT device engineering

### 3. Tiny Machine Learning

TinyML is rapidly advancing field within machine learning, specifically focused on enabling AI capabilities in resource-constrained embedded systems. It represents a transformative approach to deploying machine learning models on low-power, memory-limited devices such as microcontrollers, which are integral to the Internet of Things (IoT) ecosystem. This development allows real-time, on-device data processing and decision-making, reducing the need for constant cloud connectivity, and improving both efficiency and privacy. TinyML has wide-ranging applications across industries, including healthcare, smart cities, agriculture, and industrial automation [13],[14].

The growing adoption of TinyML underscores its potential to reshape the landscape of AI and embedded systems, offering scalable, energy-efficient solutions for real-time, intelligent decision-making across a wide range of applications.

### 4. Neuton.AI and its Features

Neuton AI is a cutting-edge platform that simplifies the development and deployment of machine learning models. Neuton AI leverages an automated machine learning approach, allowing users to create highly efficient models without requiring deep technical expertise in machine learning or data science. The platform is designed to address the challenges of deploying machine learning on resource-constrained devices, such as microcontrollers and embedded systems, by optimizing models for low memory usage, reduced computational power, and energy efficiency. With utilization of this platform, easily can be solved regression, classification and anomaly detection [7],[8].

It is commonly used in scientific and engineering fields for integrating machine learning into edge and embedded systems. The list of its diverse capabilities includes:

- **No-code model development** - a key feature of neutron.ai is its no-code environment, which allows users to develop machine learning models without the need for programming. This feature broadens access to AI development, making it more accessible to non-technical users and a wider audience.
- **Lightweight models** - neutron.ai specializes in creating ultra-compact models optimized for deployment on devices with minimal computational resources, such as microcontrollers and edge devices. These models are designed for efficient memory usage and energy consumption, making them well-suited for TinyML applications.
- **Low latency and real-time processing** - the platform emphasizes the development of models optimized for real-time execution, supporting low-latency decision-making on edge devices. This capability is essential for applications requiring immediate responses, including autonomous systems, healthcare monitoring, and predictive maintenance.



- **Data privacy and security** - by facilitating local data processing on devices rather than relying on cloud servers, neutron.ai enhances data privacy and security. This approach minimizes the transmission of sensitive information, addressing privacy concerns in fields such as healthcare and finance.
- **Cross-platform compatibility** - neutron.ai is designed for seamless integration across diverse hardware platforms and environments. Its models are deployable on a range of devices, from microcontrollers to more advanced embedded systems, ensuring flexible deployment options.
- **Energy efficiency** - its models are optimized for minimal energy consumption, which is crucial for battery-powered IoT devices and embedded systems. This efficiency extends the operational life of devices, reducing the need for frequent recharging or battery replacements.
- **Scalability** - supports scalable deployment of models across multiple devices and environments, making it ideal for large-scale IoT deployments where numerous devices require simultaneous data processing.

## 5. Utilization of Neutron.AI

The practical example of training the model we have created using data for Transport Type Detection as a source data. Next phase was training the pipeline where the model type was Multi Classification (used to predict one value of the limited number or of possible outcomes). Focus was placed on maximizing accuracy as the evaluation metric, with a training duration limited to 4 hours. The model was optimized for deployment on Intelligent Sensor Processing Units (ISPU).

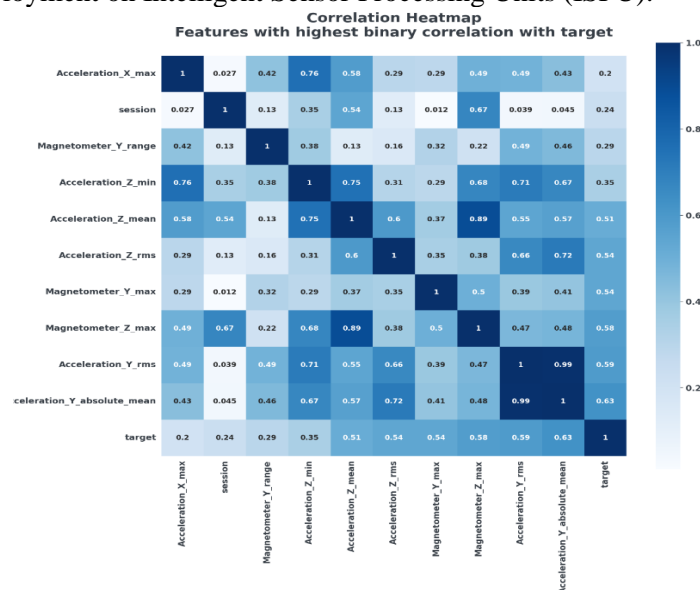


Figure 4: Heatmap of the features with highest binary correlation with target

After the phase of training the pipeline, next stage is results reviewing. Exploratory data analysis shown that for training we used data set dimensions with rows: 6210, columns:101, size memory:2.44mb. The data set based on the variables was split into 8 classes, above on the Figure 4 is shown the correlation heatmap (value 1 is indicating perfect correlation of each feature with itself, 0 indicates no correlation and -1 would indicate a perfect negative correlation, but this dataset shows only positive correlations). The top five pairs of features that exhibit high mutual correlation values exceeding 0.7, Figure 5.

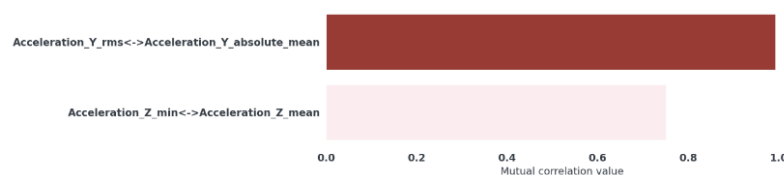


Figure 5: Columns with high mutual correlation (>0.7) TOP 5

Based on the metric type for the target hardware ISPU, there is an option to select if we want to use the holdout metrics or training metrics (more to read about the difference in Table1).

**Table 1:**  
Key differences between Training and Holdout Metrics

Feature	Training Metrics	Holdout Metrics
Data Used	Training dataset	Holdout(validation/test)dataset
Purpose	Evaluate model learning	Access generalization ability
Implications	Can indicate overfitting	Reflect real-world performance
Common metrics	Accuracy, loss, precision, recall	Accuracy, loss, precision, recall

Both training and holdout metrics are essential for building and validating machine learning models. Training metrics assess how well the model has learned from the training data, while holdout metrics evaluate the model's performance on unseen data, providing a clearer picture of its potential effectiveness in real-world applications.

In our case, with utilization of neutron.ai platform the monitoring of the holdout metrics performance for the model show 0.95% of accuracy and good performance in the FLASH and SRAM (Static Random Access Memory) memory, Figure 6.



Figure 6: Performance holdout metrics accuracy

If we focused even more on the model quality, based on the Figure 7 easy we can detect the stable version of the model.

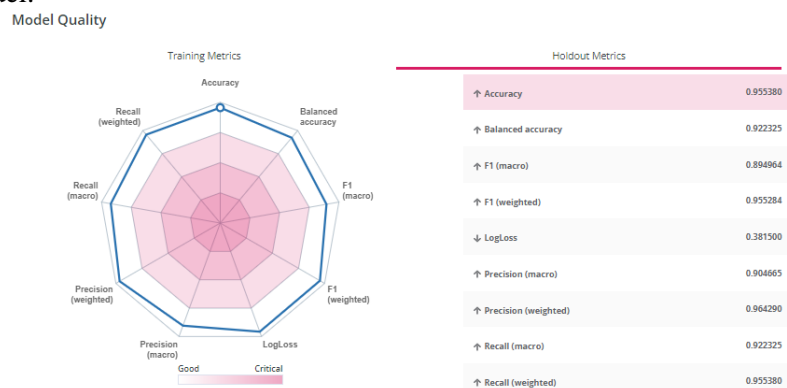


Figure 7: Analytics of the model

Classification tasks only can be represented in confusion matrix Figure 8, that shows the number of correct and incorrect predictions based on the validation data.

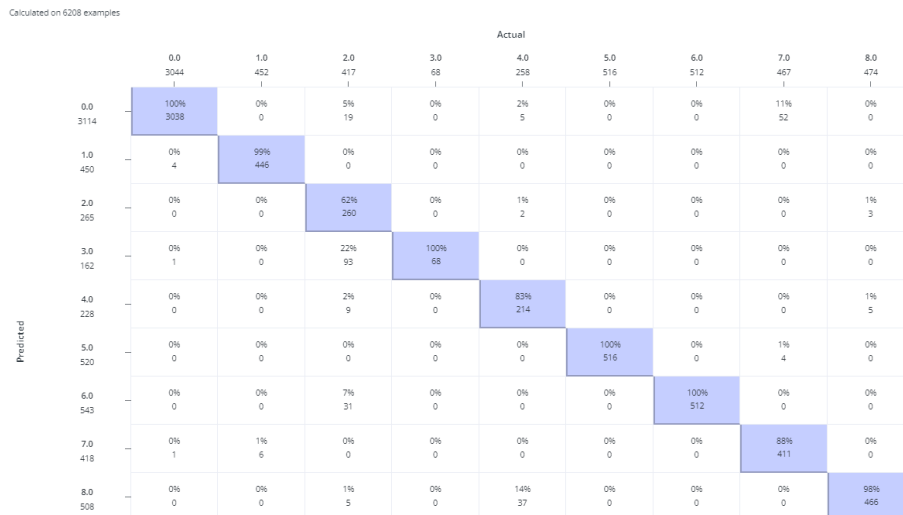


Figure 8: Confusion Matrix

After completing the training, C libraries will be prepared for a wider range of hardware types, and source code will be available for enterprise plans. With downloading the “C Library” of the selected model that can integrated it on the device Figure 9.

The “C Library” contains the following files (which is not recommended to be modified, since unsupervised changing of files can cause errors in model inference):

- **Artifacts** – contains models converted to various formats, as well as an executable file for predictions on the desktop.
- **Neuton** – supported libraries for embedding, Cortex M0, Cortex M4, Cortex M33 and STMicro ISPU.
- **Neuton-generated** – contains information necessary for the correct operation of libraries.
- **LICENSE** – contains the possibilities and restrictions on the use of its intellectual property.
- **README** – contains instructions for leveraging libraries.

Name	Type	Compressed size	Password prot...	Size	Ratio	Date modified
artifacts	File folder					10/3/2024 11:02 AM
neuton	File folder					10/6/2024 8:50 AM
neuton-generated	File folder					10/6/2024 8:49 AM
LICENSE.txt	Text Document	1 KB	No	2 KB	52%	10/3/2024 11:02 AM
README.md	Markdown Source File	4 KB	No	12 KB	73%	10/3/2024 11:02 AM

Figure 9: C Library folder structure

## 6. Future Directions and Limitation

Neuton.AI is increasingly recognized as a leading platform in TinyML and edge computing, prompting the exploration of several emerging future directions. Several future directions are emerging:

- The platform is poised to significantly enhance **AI integration at the edge** with the **expanding IoT landscape**. By supporting a wider range of edge hardware, such as microcontrollers and low-power devices, its application in areas like smart cities, autonomous vehicles and healthcare wearables can deepen human-machine collaboration and advance edge computing.
- **Enhancing real-time data processing** capabilities is crucial for meeting the growing demands of low-latency applications, such as autonomous drones and smart medical devices. Optimizing inference speed and reducing latency will be essential for advancing autonomous systems and supporting mission-critical industries.
- **On-device adaptive learning**, where models evolve from real-time data without cloud retraining, represents a key area of advancement in edge AI.

- It could expand its utility by developing **domain-specific models** tailored to fields like genomics, personalized medicine, and environmental monitoring. This would involve creating customized solutions for handling specialized datasets and optimizing algorithms for specific industry applications [15].

Despite its considerable potential, Neuton.AI encounters several limitations that require attention and resolution [16]:

- **Limited customization for advanced users** - it effectively automates model development for users with limited technical expertise. However, this focus on simplicity may restrict flexibility for experienced machine learning practitioners. Users seeking greater customization or the ability to fine-tune complex models for specific applications may find Neuton.AI automated approach insufficient for their advanced needs.
- **Challenges with ultra-low-power devices** - even highly optimized models may underperform, limiting their applicability in certain edge computing scenarios.
- **Non-Sensor-based data** - its versatility may be constrained when applied to computationally intensive tasks, such as natural language processing (NLP) or computer vision, which often necessitate larger models and greater processing power than typical edge devices can support.
- **Scalability in large-scale edge networks** - scaling across numerous edge devices introduces significant operational challenges, including model update management, device synchronization, and addressing specific hardware limitations. Although Neuton.AI performs effectively in smaller deployments, larger-scale edge AI networks may necessitate more robust infrastructure for efficient model deployment and maintenance.

## 7. Real-World Application

Today this platform is used wide world for different applications such as: predictive maintenance in industrial IoT (to monitor equipment sensors to detect unusual patterns or vibrations that indicate potential failures to reduce downtime and operational costs in industries), health monitoring devices (process physiological data like heart rate, oxygen levels or ECG patterns directly on the device without needing a cloud connection), smart agriculture sensors (detect soil moisture, temperature and crop health status), environmental monitoring (air quality), smart home and building automation (adjust lighting) and etc. Neuton.AI is adaptable across different sectors, where the platform's lightweight, cloud-independent approach makes it ideal for real-time, cost-effective AI solutions.

## 8. Conclusions

Neuton.AI presents significant advantages in enabling AI at the edge, especially through its automation and TinyML capabilities, its future success will depend on overcoming current limitations related to customization, hardware constraints, data security and scalability. As it continues to evolve, addressing these challenges will be essential to broadening its applicability and pushing the boundaries of edge AI. In our paper, we have provided a comprehensive overview of Neuton.AI's capabilities, demonstrating its platform features and evaluating its current strengths. Additionally, we identified future directions and potential innovations, particularly in embedding AI within IoT devices in more efficient and intelligent ways. These advancements promise to unlock new possibilities for real-time data processing and autonomous decision-making at the edge, further enhancing the integration of AI in smart technologies and paving the way for its expanded adoption across diverse industries.

### References:

- [1] David Hanes, Gonzalo Salgueiro, Patrick Grossetete, Rob Barton and Jerome Henry. "IoT Fundamentals Networking Technologies, Protocols and Use Cases for the Internet of things. " Cisco Press (2017): 2017937632.
- [2] Radouan A. Mouha. "Internet of Things (IoT)", Journal of Data Analysis and Information Processing (2021): 10.4236/jdaip.2021.92006

- [3] J. Gubbi, R. Buyya, S. Marusic and M. Palaniswami. "Internet of Things (IoT): A vision, architectural elements, and future directions", Elsevier - Future Generation Computer Systems (2013): <https://doi.org/10.1016/j.future.2013.01.010>
- [4] In Lee and Kyoochun Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises", Elsevier, Business Horizons Volume 58, Issue 4 pp.43-440 (2015): <https://doi.org/10.1016/j.bushor.2015.03.008>
- [5] Norah N. Alajlan and Dina M. Ibrahim. "TinyML: Enabling of Inference Deep Learning Models on Ultra-Low-Power IoT Edge Devices for AI Applications", Micromachines (2022): <https://doi.org/10.3390/mi13060851>
- [6] M. Giordano, N. Baumann, M. Crabolu, R. Fischer, G. Bellusci, M. Mango. "Design and Performance Evaluation of an Ultralow-Power Smart IoT Device With Embedded TinyML for Asset Activity Monitoring", IEEE Transactions on Instrumentation and Measurement, Volume 71, <https://ieeexplore.ieee.org/document/9758676>
- [7] C. Banbury, V. J. Reddi, A. Elium, S. Hymel, D. Tischler, D. Situnayake, C. Ward, L. Moreau, J. Plunkett, M. Kelcey, M. Baaijens, A. Grande, D. Maslov, A. Beavis, J. Jongboom and J. Quay. "Edge Impulse: An MLOps Platform for Tiny Machine Learning", Proceedings of Machine Learning and Systems 5 (MLSys 2023): [https://proceedings.mlsys.org/paper\\_files/paper/2023/file/49fe55f5e9574714dda575bfb2177662-Paper-mlsys2023.pdf](https://proceedings.mlsys.org/paper_files/paper/2023/file/49fe55f5e9574714dda575bfb2177662-Paper-mlsys2023.pdf)
- [8] Neuton.AI official documentation, <https://neuton.ai/>, last accessed 2024/09/15
- [9] Satyanarayan Kanungo. "Edge-to-Cloud Intelligence: Enhancing IoT Devices with Machine Learning and Cloud Computing ", IRE Journals, Volume 2 Issue 12 (2019): ISSN: 2456-8880
- [10] Cosmina M. Rosca. "Convergence Catalysts: Exploring the Fusion of Embedded Systems, IoT, and Artificial Intelligence", Springer Singapore (2024): [https://doi.org/10.1007/978-981-97-5979-8\\_4](https://doi.org/10.1007/978-981-97-5979-8_4)
- [11] Franklin Olivera, D.G. Costa, Flavio Assis, I. Silva. "Internet of Intelligent Things: A convergence of embedded systems, edge computing and machine learning", Internet of Things Elsevier, Volume 26 (2024): <https://doi.org/10.1016/j.iot.2024.101153>
- [12] Z. Zhang, J. Li. "A Review of Artificial Intelligence in Embedded Systems", Micromachines (2023): <https://doi.org/10.3390/mi14050897>
- [13] Syed Ali R. Zaidi, Ali M. Hayajneh, M. Hadeez, Q.Z Ahmed. "Unlocking Edge Intelligence Through Tiny Machine Learning (TinyML)", IEEE Access, vol. 10, pp. 100867-100877, (2022), doi: 10.1109/ACCESS.2022.3207200
- [14] Ji Lin, L.Zhu, Wei-Ming Chen, Wei-Chen Wang, Song Han. "Tiny Machine Learning: Progress and Futures [Feature]", IEEE Circuits and Systems Magazine, vol. 23, no. 3, pp. 8-34, (2023), doi: 10.1109/MCAS.2023.3302182.
- [15] H. B. Pasandi, F. B. Pasandi, F. Parastar, A. Moradbeikie, T. Nadeem. "Echoing the Future: On-Device Machine Learning in Next-Generation Networks – A Comprehensive Survey", ResearchGate(2023), [https://www.researchgate.net/profile/Hannaneh\\_Barahouei\\_Pasandi/publication/371139760](https://www.researchgate.net/profile/Hannaneh_Barahouei_Pasandi/publication/371139760).
- [16] A. Elhanashi, P. Dini, S. Saponara, Q. Zheng." Advancements in TinyML: Applications, Limitations, and Impact on IoT Devices", Electronics (2024): <https://doi.org/10.3390/electronics13173562> .

# Using predictive analytics and machine learning in Python to analyze and predict student success

Katarina Vignjević<sup>1</sup>, Marko Blažić<sup>1</sup>, Biljana Radulović<sup>1</sup>, Milan Marković<sup>1</sup>, Uroš Šarenac<sup>1</sup> and Dilan Dobardžić<sup>1</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

vignjevickatarina@gmail.com; markoblazic93@gmail.com; biljana.radulovic66@gmail.com; milanzrmarkovic@gmail.com; uros.sarenac@tfzr.rs; dilan.dobardzic@tfzr.rs

## Abstract:

This paper applies predictive analytics and machine learning using Python for better understanding and prediction of students' success in different educational contexts. It also discusses how predictive models, using a range of machine learning algorithms-including ensemble models-can find students at risk much earlier than is currently possible to identify the right time for appropriate interventions to retain more students and improve academic performance. Such a study, therefore, calls for comprehensive data collection on demographic, academic, and co-curricular variables to ensure the development of valid models with educative strategies and personalized academic advisement. The ethical use of predictive analytics-what has been termed as bias and transparency-implies trust among all the stakeholders concerned. As big data analytics is increasingly adopted across different educational institutions, the potentiality for changing the educational landscapes grows by leaps and bounds.

## Keywords:

Predictive analytics, machine learning, Python.

## 1. Introduction

Predictive analytics and machine learning have now combined in an approach aimed at better understanding and improving student success in the context of education. The paper describes the application of these technologies using Python in analyzing different aspects that influence students' performances and come up with predictive models that can forecast academic outcomes. This has implications for being able to provide educators and administrators with information that is actionable, leading to improved educational strategies and interventions [1].

Recent studies have emphasized the relevance of co-curricular involvement in relation to student success. A comprehensive measure of involvement provides new insights into the characteristics that predict student thriving, both quantity and quality of the experiences are found to be critical factors related to academic achievement [2]. It also talks about the implication of course modalities on student retention and success in showing that results can vary significantly depending on the demographic makeup when using different instructions as indicated by Hamann et al. 2020. This therefore, implies that in developing predictive models that are of a high degree of accuracy, there is a need to look at many sources of information [3].

The investigation of various machine learning algorithms, in a paper by Granados et al., depicted the way different types of algorithms, such as support vector machines and decision trees, could be applied to predict student performance, which has enforced the idea that machine learning can go a long way in the understanding and prediction of student success [2].

## 2. Predictive analytics in Python

Python is very popular in predictive analytics due to the versatility of the language and the huge ecosystem of libraries developed for data analysis and machine learning. This wide usage in the domain results from simple syntax and a set of powerful libraries such as Scikit-learn, TensorFlow, Pandas,

matplotlib, Sklearn and NumPy that assure ease in development and enhancement of the predictive model across various applications like healthcare, finance, and transportation [4].

One of the most striking features of predictive analytics is its modeling capability and classifying and predicting results using data recorded previously. For example, Geldiev et al. identify that the whole purpose of the training process is to come up with a model that has actually been developed from classified examples, after which one can classify any new data points using the same model quite effectively. This iterative process of model training and validation is needed, as various studies done on Python libraries for tasks ranging from predicting diabetes to analyzing employee attrition [5][6].

Transportation has also applied it in many ways, such as optimizing intelligent transport systems. One such example is that Xiao in his paper presents the methods of using different Python libraries like Pandas and SciPy for analysis of traffic flow data and forecast traffic flow to improve safety and efficiency in the use of roads within urban mobility. Such examples point toward Python's ability to adapt according to very different aspects of predictive analytics challenges [7].

Besides, the development of dedicated libraries like Maestro serving to optimize energy networks and CAFA-evaluator, developed to benchmark classification methods, serves to show that not all innovation stops within the Python ecosystem. The tools simplify not just the process of predictive modeling but also make various advanced analytical techniques more accessible to practitioners in different types of science fields [8].

### 3. Machine learning in Python

Machine learning has now turned into an important component of modern data analysis and artificial intelligence. Python is turning out to be the most used programming language in this area. The reason of its growing popularity is the availability of a great number of libraries and frameworks, which allow easy implementation of all sorts of machine learning algorithms. Python becomes a critical tool for both researchers and practitioners, as libraries like Scikit-learn, TensorFlow, and PyTorch allow one to carry out both classic and deep learning.

Python in Machine Learning is versatile in that it covers quite a great number of algorithms and methodologies. For example, Scikit-learn provides a broad range of algorithms for classification, regression, and clustering under a uniform interface that facilitates the modeling work in a comfortable way. On the other hand, TensorFlow and PyTorch are used for deep learning applications and provide high-class capabilities to build and train neural networks. This rich ecosystem not only enhances productivity but also allows for rapid prototyping and experimentation, which is imperative in the fast-evolving field of machine learning [8][9].

Python also works well with other libraries that manipulate data, such as NumPy and Pandas. Both libraries enable efficient handling and preprocessing of data, critical steps in a machine learning pipeline. It is a massive advantage because big datasets will be processed with much ease; hence, the researcher can focus on model development rather than data management challenges. The syntax in Python is relatively friendly to users, making the language accessible to newcomers in the field and hence widening the machine learning practitioners' community [4][8][9].

Machine learning application in Python is materialized for many fields: healthcare, finance, environmental science, and so on. For example, research has illustrated the use of Python for the drug classification task in which machine learning techniques were used to increase the accuracy of the prediction. Similarly, Python has been implemented in the analysis of environmental data further showing how it can solve real-world problems using machine learning. This does support further versatility with increased educational resources and community contributions that allow it to enhance learning and collaboration while working in the Python ecosystem. [9]

### 4. Predictive analytics and machine learning in education

Predictive analytics and ML for education have, over the years, really taken off and have been major aids in analysis and predictions concerning the success of students. Such techniques make use of historical data to define trends and influencing factors concerning academic success; hence, they allow for targeted interventions to be carried out by the respective institutions.

The first major benefit of the usage of predictive analytics in education is objective analysis of students' performance facilitated through various types of machine learning models. For instance, it was revealed that models such as RF may provide a pretty good forecast of students' academic achievements according to their physical activity and level of stress. These models not only bring insights into major predictors of success, but also allow on-time strategic changes in teaching strategies depending on the forecasted results. Besides, current literature and practice show that integrated deep learning methods promise high-level feature extraction in complex data, which in turn offer improved predictive accuracy [4][10].

Secondly, it is of utmost importance to have high-quality and comprehensive data. In fact, any predictive modeling requires the availability of pertinent and accurate data on demographic, academic, and even extracurricular details of students. The successful implementation of these models is contingent upon addressing the issues of data privacy and security. Sensitive information is to be guarded but still be used for improvements in education. Feature selection is very important, as it directly influences the performance of the model and thereby the insights derived from the model [4][10].

Comparative studies of various predictive models have also comprised an important part of recent work in this area. Several reports have suggested that a systematic comparison of various approaches to modeling is required in order to ascertain the relative merits and limitations of each. Classification algorithms like ANN, SVM, and Decision Trees have been used many times to classify students' performance as pass/fail or excellent/poor. These kinds of comparison frameworks provide educators with assistance in making relevant decisions so as to enhance the outcomes of their students [4][10].

Further, influences have been there from socio-demographic parameters in predicting the success of students. This is underpinned by the fact that factors such as economic background, previous academic success, and academic activities engagement are strong predictors of students' adaptation and achievement in colleges. With the consideration of these factors in predictive models, by all indications, the adjustments of support responsiveness will also be more sensitive to individual needs within their diverse demographics [10][11][12].

#### **4.1. Predictive analytics and Python for students success prediction**

Predictive analytics with Python in Student Information Systems has recently drawn great attention when it comes to the analysis and predictions for better academic success. The approach shall make use of the large volume of data generated by students to glean insight and make informed predictions about academic performances that might help improvement in student outcomes [13].

Predictive analytics makes use of statistical methods and machine learning algorithms to analyze data from the past in order to predict what is most likely to happen in the future. Such analytics could include demographic data, academic success or performance, attendance history, and various measures related to student engagement in an attempt to predict the likelihood of successful or failing students [13].

For instance, Chen and Upah explain how predictive analytics can be applied to academic advising to provide feedback that is data-driven, which in turn encourages students to engage in sense-making activities with respect to their academic trajectories, and it proves that learning in the perspective of data analytics informs decisions on how institutions can be certain about student retention and success. This ability to preprocess and analyze, using Python libraries such as Pandas and Scikit-learn, would, in turn, enable developers to create robust predictive models that could work across multiple educational contexts [13].

Furthermore, the integration of machine learning with big data analytics provides deeper insight into the behaviors and learning patterns of students. Big data may explain the learning needs of students and, therefore, facilitate the implementation of personalized interventions to achieve better academic performance. This view has been taken up by Bai, 2024: "Big data can help explain the students' learning needs, thus facilitating the implementation of personalized interventions to achieve better academic outcomes". This is in agreement with the findings of Hall et al., in which they established that proactive student-success coaching-informed predictive analytics significantly improve academic performance and persistence among community college students [14][15].



The general steps in any predictive modeling, therefore, include data collection, preprocessing, model selection, training, and evaluation. Such is the case where Cui et al. mention that learning analytics is starting to be one of those areas where institutions try to improve student experiences by developing predictive models based on student data. Because the chosen algorithms, such as decision trees, support vector machines, or neural networks, determine significantly the effectiveness of the prediction which could be done, besides being interpretable; for example, according to Bird et al., one way to mitigate bias in predictive models to achieve unbiased assessment of students is the removal of demographic predictors [16][17].

## 5. Predictive model example

In the present educational landscape, one of the important tasks that educators and institutions can undertake is to predict student success. Using data analysis and machine learning techniques, schools can detect early at-risk students and intervene with appropriate support to improve their academic performance. This is usually done by building a predictive model that can use various types of student characteristics, such as study hours, attendance, and previous academic performance, to predict their chances of success [18].

López et al. (2013) conducted a systematic review, "Predicting students' academic performance in higher education," reviewing the myriad of studies on forecasting academic outcomes using different predictive modeling techniques. The authors first outline the increasing importance of data-driven approaches in educational settings and highlight the role such models may play in identifying at-risk students and customizing interventions to secure their success. Based on the findings of several research works, this review aggregates methodologies adopted for analysis, which have ranged from classic statistical techniques to advanced machine learning algorithms.

One central insight that arises from the review is the various influences which might predict students' performance. López et al. group these predictors into three broad categories: demographic factors, including variables related to age, gender, and socio-economic status; academic factors, including previous grades and study habits; and psychological factors, such as motivation and self-efficacy. The multi-faceted approach underlines the complexity of academic success and the necessity to consider various dimensions within predictive modeling. One more section of the review presents the efficiency of different techniques of modeling, according to the data, a machine learning approach and specifically decision trees and neural networks are promising for enhancing accuracy within prediction.

However, the authors also raise awareness about the ethical issues of predictive models in education. They support responsible use of data to ensure that interventions driven by predictive analytics will not further entrench existing inequities or biases. By supporting transparency and fairness in model development, López et al. also emphasize the collaboration needed between educators, data scientists, and policymakers. While this systematic review underlines progress in predictive modeling for academic performance, it also challenges a more nuanced understanding of how these tools can be applied to foster equitable educational outcomes [19].

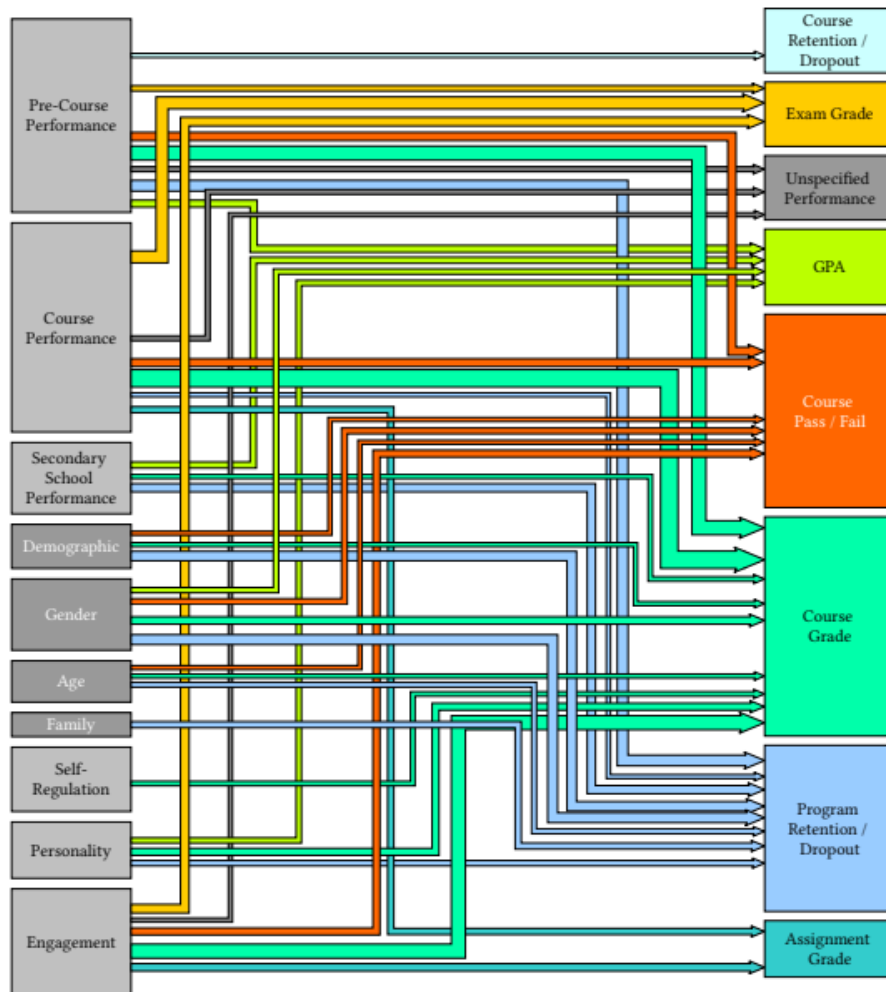


Figure 1: Most frequently researched features as predictors (left side) for predicted values (right side) [20].

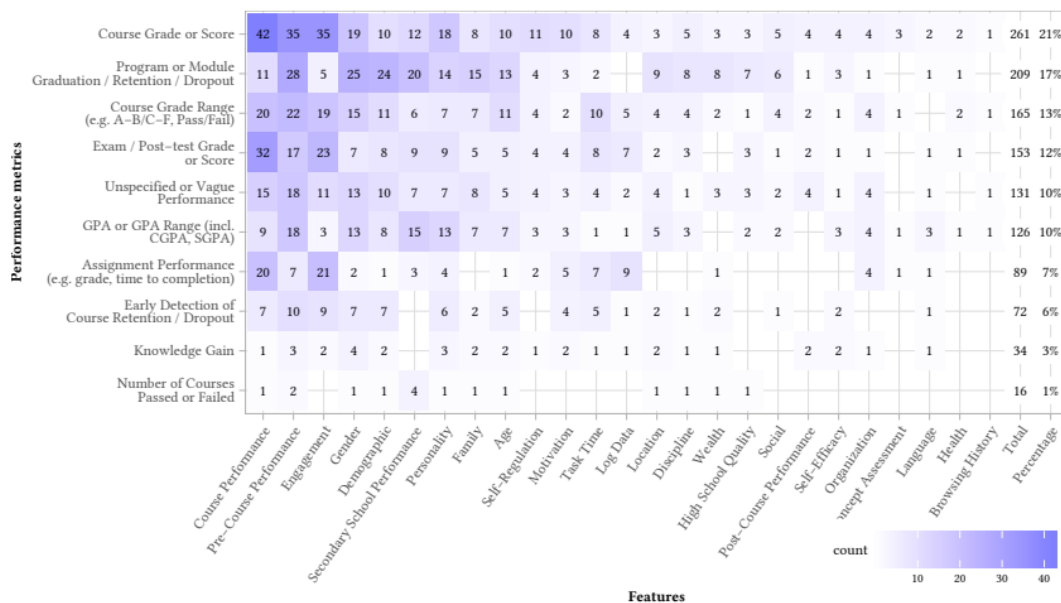


Figure 2: Cross-tabulation of features (inputs) and performance values (outputs) [20].

## 6. Conclusions

In conclusion, the introduction of predictive analytics coupled with Python machine learning makes all the difference in educational methods by showing early identification of at-risk students and informing timely interventions toward the better retention and success of all students. They can, therefore, apply different machine learning techniques, such as ensemble models, to understand performance patterns and redesign academic advising for individual learning trajectories.

But what is very important now is that in advancing these technologies, issues of bias and transparency take center stage to gain the trust of both students and stakeholders. As great data and learning analytics continue to be adopted by educational institutions, it holds a great prospect for transforming the educational landscape.

It is further promised that deep learning, among other analytical frameworks, will result in finer-grained understandings of student behavior, better teaching strategies, and improved educational outcomes. All this convergence of data-driven decision-making and educational practices heralds a new era where student success and institutional effectiveness could be significantly enhanced.

### References:

- [1] Bakker, T., Krabbendam, L., Bhulai, S., Meeter, M., & Begeer, S. (2023). Predicting academic success of autistic students in higher education. *Autism*, 27(6), 1803-1816. <https://doi.org/10.1177/13623613221146439>
- [2] Granados, D., Ugalde, J., Salas, R., Torres, R., & López-Gonzales, J. (2022). Visual-predictive data analysis approach for the academic performance of students from a peruvian university. *Applied Sciences*, 12(21), 11251. <https://doi.org/10.3390/app122111251>
- [3] Hamann, K., Glazier, R., Wilson, B., & Pollock, P. (2020). Online teaching, student success, and retention in political science courses. *European Political Science*, 20(3), 427-439. <https://doi.org/10.1057/s41304-020-00282-x>
- [4] Vignjević, K. (2024). Primena alata poslovne inteligencije u analizi uspešnosti studiranja na fakultetima. Master rad. Tehnički fakultet "Mihajlo Pupin". Zrenjanin. 20-22.
- [5] Geldiev, E., Nenkov, N., & Petrova, M. (2018). Exercise of machine learning using some python tools and techniques. *Cbu International Conference Proceedings*, 6, 1062-1070. <https://doi.org/10.12955/cbup.v6.1295>
- [6] Gorecki, T. and Martin, W. (2019). Maestro: a python library for multi-carrier energy district optimal control design.. <https://doi.org/10.48550/arxiv.1911.12661>
- [7] Xiao, K. (2024). Enhancing urban mobility: a comprehensive evaluation of python's impact in optimizing intelligent transportation systems.. <https://doi.org/10.3233/atde24037>
- [8] Albertsson, K., An, S., Gleyzer, S., Moneta, L., Niermann, J., Wunsch, S., ... & Mesa, O. (2020). Machine learning with root/tmva. *Epj Web of Conferences*, 245, 06019. <https://doi.org/10.1051/epjconf/202024506019>
- [9] Alnaasan, N., Jain, A., Shafi, A., Subramoni, H., & Panda, D. (2021). Omb-py: python micro-benchmarks for evaluating performance of mpi libraries on hpc systems.. <https://doi.org/10.48550/arxiv.2110.10659>
- [10] Akour, M., Sghaier, H., & Qasem, O. (2020). The effectiveness of using deep learning algorithms in predicting students achievements. *Indonesian Journal of Electrical Engineering and Computer Science*, 19(1), 388. <https://doi.org/10.11591/ijeecs.v19.i1.pp388-394>
- [11] Arora, D. (2023). Comprehensive analysis of factors influencing the real-world application of machine learning for student success rate calculation and their impacts on student achievement & educational institutions. *World Journal of Advanced Research and Reviews*, 19(3), 942-953. <https://doi.org/10.30574/wjarr.2023.19.3.1879>
- [12] Byshevets, N. (2024). The influence of physical activity on stress-associated conditions in higher education students. *Teoriâ Ta Metodika Fizičnogo Vihovannâ*, 24(2), 245-253. <https://doi.org/10.17309/tmfv.2024.2.08>
- [13] Chen, Y. (2024). A comparative study on the results of college english grade 4 based on multi-model prediction. *jes*, 20(6s), 387-392. <https://doi.org/10.52783/jes.2660>

- [14] Hall, M., Worsham, R., & Reavis, G. (2021). The effects of offering proactive student-success coaching on community college students' academic performance and persistence. *Community College Review*, 49(2), 202-237. <https://doi.org/10.1177/0091552120982030>
- [15] Bai, H. (2024). Design and application of decision support system for educational management based on big data. *jes*, 20(6s), 1645-1655. <https://doi.org/10.52783/jes.3084>
- [16] Cui, Y., Fu, C., Shiri, A., & Yang, F. (2019). Predictive analytic models of student success in higher education. *Information and Learning Sciences*, 120(3/4), 208-227. <https://doi.org/10.1108/ils-10-2018-0104>
- [17] Bird, K., Castleman, B., Mabel, Z., & Song, Y. (2021). Bringing transparency to predictive analytics: a systematic comparison of predictive modeling methods in higher education. *Aera Open*, 7. <https://doi.org/10.1177/23328584211037630>
- [18] Baker, R. S. J. d., & Inventado, P. S. (2014). "Educational Data Mining and Learning Analytics." In *Learning, Design, and Technology: An Encyclopedia for Teachers*. Springer.
- [19] López, M. C., et al. (2013). "Predicting students' academic performance in higher education: A systematic review." *Computers & Education*, 58(1), 13-20.

# An Overview of Multimedia - audio data visualization process using Business intelligence tools

Marko Blažić<sup>1,2</sup>, Srđan Popov<sup>2</sup>, Višnja Ognjenović<sup>1</sup>, Katarina Vignjević<sup>1</sup>, Milan Marković<sup>1</sup>, Slobodan Nadrljanski<sup>1</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin

<sup>2</sup> University of Novi Sad, Faculty of Technical Science, Novi Sad

marko.blazic@tfzr.rs; srdjanpopov@uns.ac.rs; visnjaognjenovic@gmail.com;  
katarina.vignjevic@tfzr.rs; milan.markovic@tfzr.rs; sloba92ki@gmail.com

## Abstract:

This synthesis examines the critical role of data visualization in enhancing insights and decision-making across various sectors, particularly in business intelligence (BI) and audio analytics. It highlights the necessity of selecting appropriate visualization techniques—such as bar charts, line charts, and heat maps—tailored to the characteristics of the data and the familiarity of the audience. The text emphasizes that the effectiveness of visualizations is not just about aesthetics but also about improving understanding and facilitating informed decisions. In the realm of BI, interactive visualizations allow users to dynamically explore complex datasets, fostering a deeper comprehension of underlying trends and reducing cognitive overload. Furthermore, the dynamic integration of audio data visualization within BI frameworks, particularly through tools like Microsoft Power BI, illustrates how organizations can leverage audio datasets for behavioral monitoring and emotional analysis. The synthesis discusses advancements in audio-visual integration that enhance tasks like emotion recognition, making data analysis more robust and accurate. Moreover, it explores how techniques like compression optimize audio data for visualization. Ultimately, the synthesis underscores the importance of thoughtful visualization strategies in turning data into actionable insights, providing organizations with a strategic advantage and enhancing overall performance.

## Keywords:

Visualization process, Visualization data types, audio data, audio data in visualization process

## 1. Introduction

Humans are visual beings from early ages. From hunting strategies to all kinds of animal they have encountered, it is possible to conclude that this were the moment of necessity for data visualization. Many of ancient human letters are visual based, civilization like Egyptians and Mayas created photographic images, that allows communication between people and through generations. Maps are also used to show the wealth of the civilization [1].

The map of the city of Ga Sur is the most antique map, and it dates from 2500 B.C., and presents Euphrates river between two mountains. Also, there are Homero map and Ptolemy map, two maps very important for the ancient world. In sixteenth century, Portuguese wealth was presented as shown in the Cantino map [1].

William Playfair wrote a book with diverse types of diagrams to show all kinds of statistic information. In the seventh decade of last century, the first infographic appeared in magazines and paper, in order to gather information and present them graphically, in the best possible way for that kind of information [1].

Nowdays, data is everywhere around us. As a result of increased amount of generated data, the need for data visualization is growing. Companies are generating large amounts of data everyday and as a result, data available on web has been increasing drastically. It is very difficult for users to analyze and explore large quantities of data, so data visualization comes useful [2].

Data visualization is based on things such as development, design and graphical presentation of data. This action provides ability for customers to analyze data through visual charts, that originates from different sources. Visual charts are also good for discovering data patterns, and forming a conclusion [2].

Data visualization is also known as a scientific visualization or information visualization. The main goal of visualization is to make information last through time. All information that are not possible to be tasted, touched or smelled, have to be interpreted visually [3].

## 2. Visualization data types

Data visualization is an important aspect of data analysis in the sense that it gives users some insight into a rather complex array of data visualized graphically. How various types of visualization work out best depends on such factors as the nature of the data under consideration, audience familiarity with the visualization techniques, and the goals of the analysis, among other things. This synthesis discusses the types of data visualizations available, their applications, and considerations that ought to be in place if good implementation of the data is to occur.

There are different types of visualizations that exist for purposes and suit various characteristics of the data. Common types of visualizations include bar charts, pie charts, line charts, and heat maps, which have different advantages depending on the data being represented. For example, bar charts provide the best method of comparison in data that is categorical, while line charts are best suited for the indication of trend over time. On the other hand, the choice of visualization could make all the difference in audience understanding and decision-making. A number of studies show that user preferences and familiarity with particular kinds of visualizations can lead to misalignments with best practices in data visualization [4] [5].

In domains such as healthcare, the design of visualizations has to consider manifold needs of practitioners as well as other stakeholders. The work of Backonja et al. underlines that personal and user-centered design can enhance usability in visualizations among healthcare providers [6]. This is particularly apropos in contexts where data visualization supports decision-making, as observed in the preferred types of visualization among health professionals in the study on Antimicrobial Stewardship by Keizer [5]. Their effectiveness depends on the clear and efficient transmission of information that enables one to derive meaningful insights from the data interpreted.

Lastly, it might be higher-order visualizations involving more advanced ways of showing information; interaction and dynamic change could also be associated with it. In the words of Kharakhash [7], effective visualization turns data analysis from an exercise in technicality into a strong tool for deriving actionable insight. This is especially important in the big data environment in which such schemas are hard to represent using traditional ways of visualization in order to understand the large and complex nature of the data sets involved. Automation of the visualization process itself through the development of frameworks that can automatically decide what type of visualization to use and how to present the data, depending on user objectives, is also touched on by Golfarelli and Rizzi [8].

Data type visualization is a complex activity that involves considerations of data characteristics, user preferences, and use context. By employing adequate visualization techniques and learning from the needs of one's audience, it is possible for practitioners to make the communication of data more effective and clear

## 3. The Visualization Process in Business Intelligence

Business intelligence tools use visualization processes, which remain an essential component for improving decision-making capabilities across a wide array of sectors. Visualization in BI systems enables organizations to transform complicated datasets into easily understandable visual patterns that can be more easily analyzed for insights and interpretation. Such processes are particularly important within big data contexts since the scale and intricacy of the information could overwhelm traditional analytic methodologies [9].

However, interactive data visualization has become one of the most powerful tools in BI, letting users dynamically explore information and gain insights that might have been obscured by static reporting. Research has demonstrated that interactive visualizations greatly enhance decision-making processes by allowing users to change some dimensions of the data and observe various scenarios. This is core to organizations trying to get strategic advantage from data, since it fosters deeper understanding of underlying trends and patterns [9].

Additionally, there are many industries where the usage of BI tools exposes different facets of data visualization. For instance, in the retail industry, Power BI has been one of the BI tools used to effectively analyze sales trends. By visualizing the trend of past sales, businesses can identify patterns in growth and therefore make informed decisions on how much inventory to maintain or how effective their marketing strategy should be. Similarly, in the realm of government data, extraction and interpretation of emerging topics from open government data have utilized visualization tools and enhanced competitive intelligence accordingly [10] [11].

Key role of BI also pertains to visualization for its functionality in reducing cognitive overload originating from data analysis. BI tools use information visualization through different dashboards and graphs that allow users to process and remember the information received more easily. This is very important in daily environments with demands for speedy decisions, as it provides a possibility for stakeholders to understand even the most complex information within a very short time and with complete accuracy [12].

Another good example of how BI tools can be used to improve organizational performance is through the development of customized dashboards that combine consumer information emanating from various sources. Such custom dashboards make visualization of data easier and also allow embedding of real-time analytics, thus enabling firms to respond more quickly against fluctuating market conditions. The application of visible models in representing KPIs, therefore, facilitates strategic management in the form of a clear understanding of the metrics of organizational performance [13].

#### **4. Audio data in visualization**

Audio and visual data integration has become a key task in various applications; it thus plays a crucial role in enhancing the understanding and processing of multimodal information. Such synthesis is notably relevant to applications of audio-visual speech recognition, emotion analysis, and event localization, whereby its endpoint might be an improved performance together with robustness due to combined auditory and visual cues [14].

One of the main benefits of audio-visual data is increasing speech recognition; its growth is weakened due to the presence of noise. It has, in fact, been proven through research that the addition of visual features into speech can improve the rate of recognition in conditions where audio signals are at a loss. The results had concluded that the visual cues such as lip movements identify speech when the audio is not clear and proper. This phenomenon is termed a "visual Lombard effect" because the loss in auditory signals is restored through the visual information and hence, one understands speech better. Furthermore, the integration of audio and visual signals allows for the development of more sophisticated models that can leverage both modalities to achieve higher accuracy in speech recognition tasks [15].

In addition to speech recognition, audio-visual data plays a crucial role in emotion analysis. The ability to analyze emotions from audio-visual data offers significant advantages over using audio or video alone. In fact, studies show that auditory and visual signals add more depth to specifying emotional states, and each has a unique contribution to the perception of emotions. These have been employed in creating synchronized facial expressions with vocal intonations to enable more accurate emotion recognition, extremely useful in human-computer interaction and social robotics. With that, there is even greater need for a model that will be able to learn audio-visual data effectively, an area where machine learning techniques have also seen much improvement. Self-supervised learning methodologies have also been utilized to present a better representation of audio against the aligned visual data for developing the contrastive model that will be able to differentiate different events related to audio-visual. Methods applied support correspondence in audio-visual and provide better performance in the downstream tasks such as sound localization and event detection [16].

Another important factor is that the challenge in the use of such audio-visual data lies in the process of audio-visual synchronization. In fact, studies that have been done on this topic have indeed established that there is a window of tolerance for asynchrony between the audio and visual signals, which can affect event perception. Any application that relies on the precision of timing coherence between an audio and a visual input, such as multimedia content and virtual reality environments, relies for its functionality on such issues of synchronization being understood and addressed [17].

## 5. Audio Data Visualization in Microsoft Power BI

Audio data visualization in the Microsoft Power BI environment is one of the most important innovations in the integration of audio analytics into business intelligence frameworks. In particular, this multi-functional tool empowers users to create dynamic visualizations of audio data, which would allow a greater comprehension of their practical use in behavioral monitoring and emotional recognition.

Another key use of Power BI in audio data visualization pertains to processing and representing complex data sets that are also derived from audio. For example, Wark discusses the use of Power BI at the Dallas Zoo regarding elephant behavior. The same is gathered using RFID sensors and visualized in real time. Application to power: this is the real deal of how Power BI can ensure timely responses to changes in behavior through daily, visually intuitive reports to the managers. This visual could even be extended by means of audio data, hence enriching the insights that come from behavioral analytics to develop an understanding of animal welfare [18].

Also, the quality of data to be visualized in Power BI can be enhanced using advanced audio processing. Techniques like compression through DWT and RLE can be utilized in order to optimize the audio data for analysis. Due to this compression, a similar set of data can be handled in a much more efficient manner by organizations that can thereafter be represented on Power BI dashboards without significant loss of quality. This will help not only in better data handling but will also ensure that the visualizations are apt and reflect the right and exact audio information [19].

Audio data to recognize human emotions: Audio data can be used for the extraction and visualization of insights on human emotions by speech patterns. The following examples are frameworks that are based on MSCNN for feature extraction from audio, effectively, in emotion recognition tasks. These can then be further combined into Power BI with the intent of creating visuals that expose emotional trends over time, offering very useful insights in areas as varied as customer service or even mental health monitoring [20].

Again, the integration of audio and visual data can further enhance the analytical capability of Power BI. Researchers in the merger of audio features and visual data have proved that a merger can further enhance accuracy in tasks such as the classification of an audio-visual scene. Here, the multimodal approach offers a more enriched dataset, which can be visualized through the application of Power BI; hence, the user has the ability to see patterns and relationships that they could not have considered when working with independent audio data [21].

## 6. Real life example

A media company is interested in understanding the performance of its podcast episodes in terms of listener engagement. Each podcast episode is made up of audio files with metadata in various forms, including duration, speech-to-text transcripts, speaker identification, and emotional tone through the use of audio sentiment analysis tools.

### 6.1. Application of Audio data in Power BI

#### 1. Transcription and Speech to Text Data:

- The media company makes use of speech-to-text software. It converts the audio data of the podcasts into text. Transcript data consists of time-stamped words and can reveal over time the topics of conversations.



- Power BI Visualization:
  - Line chart or Area chart: This chart can be used to depict what topics or keywords were discussed at any single point in time throughout the podcast. For example, it can reveal how often certain keywords come up in timeline view and how long each discussion lasts.
- 2. Audio Sentiment Analysis:
  - It uses an audio sentiment analysis tool that processes the underlying audio data to provide sentiment scores-positive, negative, and neutral-based on the tone of the voice of the speaker during every moment of the podcast.
  - Power BI Visualization:
    - A stacked bar chart or heat map could give the distribution of sentiment throughout the podcast. For example, the Y-axis could represent sentiment: positive, neutral, negative; this would be plotted on the X-axis for different time segments in a podcast and give insight into how the audience or speakers' emotional tone changed in key moments of the show.
- 3. Speaker Identification & Engagement:
  - Processing of audio data in context, or in relation to the current speakers, such as hosts or guests, and their respective speaking times. This data can also capture listener engagement in terms of which parts were paused, hesitated, or listened to at a faster speed.
  - Power BI Visualization:
    - This is visualized in a pie chart format for the episode by showing the percentage of time that each speaker was speaking. A line chart visualizes the speaking pace over time-when it went up or down. This would then be matched with listener engagement metrics, such as peaks in playback or social media activity around those specific points in the episode.
- 4. Listener Engagement:
  - The podcast platform also tracks when listeners skip, rewind, or stop the episode, which is correlated with the audio file's timeline.
  - Power BI Visualization:
    - Graph listener drop-off points using a line chart or scatter plot. Then, overlay this information with the audio data-for example, any point where there was too long of a pause or an off-topic discussion-to gauge what makes listeners disengage.

#### Data Workflow:

Audio files are processed, such as speech-to-text, sentiment analysis, and speaker recognition, outside of this environment. The data imports into Power BI as structured tables, including timestamp information, speakers information, and sentiment scores. Power BI visualizes metadata in a way that can give views to the media company in intuitive, powerful, and visual ways about the behaviors of both listeners and speakers.

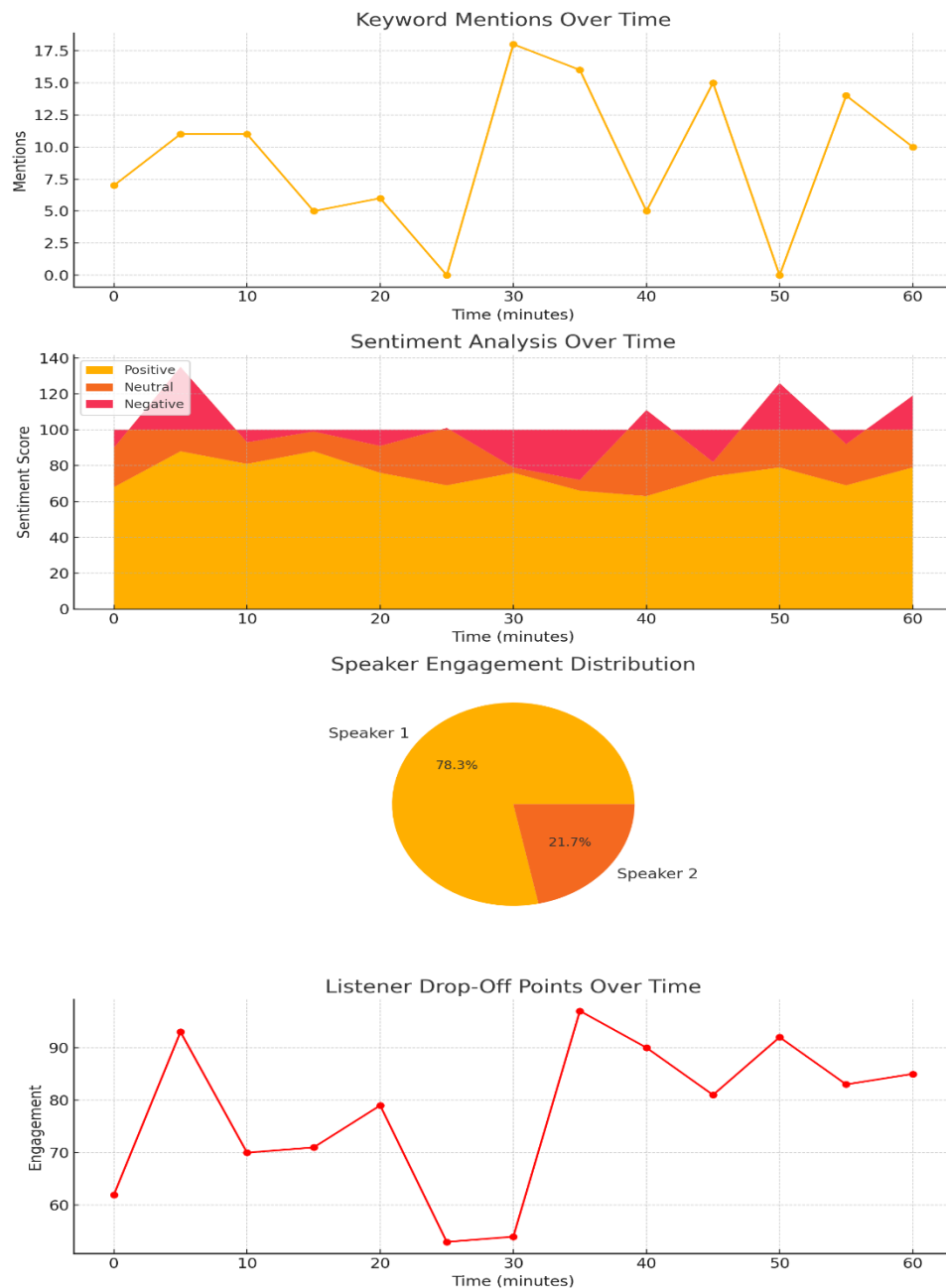


Figure 1: Podcast Performance Analysis: Audio and Listener Insights

From the previous figure we can see next:

1. **Keyword Mentions Over Time** - This shows the number of times certain keywords are mentioned throughout the podcast at specific time intervals.
2. **Sentiment Analysis Over Time** - The following stacked area chart presents the distribution of positive, neutral, and negative sentiment scores across different time segments.
3. **Speaker Engagement Distribution** - This is a pie chart showing contribution in terms of time for each speaker throughout the podcast.
4. **Listener Dropoff Points Over Time** - This graph shows the points of time at which listeners drop off, which gives some insight into listener behavior during the episode.

## 7. Conclusions

Data visualization is essential for enhancing understanding and decision-making across various domains by presenting complex datasets in accessible formats tailored to audience needs. Interactive visualization tools, particularly in business intelligence, allow users to dynamically explore and extract insights from big data, overcoming limitations of traditional reporting. The integration of audio data visualization, as seen in platforms like Microsoft Power BI, further enriches analytical capabilities by merging auditory and visual information to improve behavioral monitoring and emotional recognition. Ultimately, effective data visualization techniques, informed by user requirements and data characteristics, are crucial for clear and impactful communication of insights, driving informed decisions across diverse sectors. At the end of paper, the example of the usage of audio data in visualization process has been given, in real life example of analyzing one episode of audio podcast. A figure is presented with specific parameters which are described in previous application and for every parametar is usability and functionality is described.

### References:

- [1] Manuela Aparicio and Carlos J. Costa. 2015. Data visualization. *Commun. Des. Q. Rev* 3, 1 (November 2014), 7–11. <https://doi.org/10.1145/2721882.2721883>
- [2] Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M., & Perry, R. (2016). Data visualization. *International Journal of Engineering Research And Advanced Technology (IJERAT)*, 2(12), 11-16.
- [3] J. L. V. Sancho, J. C. Dominguez, and B. E. M. Ochoa, “An approach to the taxonomy of data visualization,” *Revista Latina de Comunicacion Social*, vol. 69, 2014, pp. 486-507
- [4] Aung, T., Niyeha, D., Shagihilu, S., Mpembeni, R., Kaganda, J., Sheffel, A., ... & Heidkamp, R. (2019). Optimizing data visualization for reproductive, maternal, newborn, child health, and nutrition (rmnch&#x26;n) policymaking: data visualization preferences and interpretation capacity among decision-makers in tanzania. *Global Health Research and Policy*, 4(1). <https://doi.org/10.1186/s41256-019-0095-1>
- [5] Keizer, J., Luz, C. F., Sinha, B., Gemert-Pijnen, L. v., Albers, C. J., Jong, N. B., ... & Glasner, C. (2021). The visual dictionary of antimicrobial stewardship, infection control, and institutional surveillance.. <https://doi.org/10.1101/2021.05.19.444819>
- [6] Backonja, U., Haynes, S. C., & Kim, K. (2018). Data visualizations to support health practitioners' provision of personalized care for patients with cancer and multiple chronic conditions: user-centered design study. *JMIR Human Factors*, 5(4), e11826. <https://doi.org/10.2196/11826>
- [7] Kharakhash, O. (2023). Data visualization: transforming complex data into actionable insights. *Automation of Technological and Business Processes*, 15(2), 4-12. <https://doi.org/10.15673/atbp.v15i2.2520>
- [8] Golfarelli, M. and Rizzi, S. (2019). A model-driven approach to automate data visualization in big data analytics. *Information Visualization*, 19(1), 24-47. <https://doi.org/10.1177/1473871619858933>
- [9] Zhang, Q. (2024). The impact of interactive data visualization on decision-making in business intelligence. *Advances in Economics, Management and Political Sciences*, 87(1), 166-171. <https://doi.org/10.54254/2754-1169/87/20241056>
- [10] Anardani, S., Azis, M. N. L., & Asyhari, M. Y. (2023). The implementation of business intelligence to analyze sales trends in the indofishing online store using power bi. *Brilliance: Research of Artificial Intelligence*, 3(2), 300-305. <https://doi.org/10.47709/brilliance.v3i2.3232>
- [11] Gottfried, A., Hartmann, C., & Yates, D. A. (2021). Mining open government data for business intelligence using data visualization: a two-industry case study. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(4), 1042-1065. <https://doi.org/10.3390/jtaer16040059>
- [12] Nwinyokpugi, P. N. (2021). Business intelligence applications; sustaining retail businesses in rivers state, nigeria. *International Journal of Multidisciplinary Research and Analysis*, 04(03). <https://doi.org/10.47191/ijmra/v4-i3-13>

- [13] Kolychev, V. D. and Shebotinov, A. (2019). Application of business intelligence instrumental tools for visualization of key performance indicators of an enterprise in telecommunications. *Scientific Visualization*, 11(1). <https://doi.org/10.26583/sv.11.1.03>
- [14] Cornu, T. L. and Milner, B. (2015). Reconstructing intelligible audio speech from visual speech features. *Interspeech 2015*. <https://doi.org/10.21437/interspeech.2015-139>
- [15] Michelsanti, D., Tan, Z., Zhang, S., Xu, Y., Yu, M., & Yu, D. (2021). An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 1368-1396. <https://doi.org/10.1109/taslp.2021.3066303>
- [16] Kadiri, S. R. and Alku, P. (2022). Subjective evaluation of basic emotions from audio-visual data. *Sensors*, 22(13), 4931. <https://doi.org/10.3390/s22134931>
- [17] Lee, J. and Ebrahimi, T. (2011). Audio-visual synchronization recovery in multimedia content. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). <https://doi.org/10.1109/icassp.2011.5946937>
- [18] Wark, J. D. (2022). Power up: combining behavior monitoring software with business intelligence tools to enhance proactive animal welfare reporting. *Animals*, 12(13), 1606. <https://doi.org/10.3390/ani12131606>
- [19] Tsegaye, A. and Tariku, G. (2019). Audio compression using dwt and rle techniques. *American Journal of Electrical and Electronic Engineering*, 7(1), 14-17. <https://doi.org/10.12691/ajeec-7-1-3>
- [20] Peng, Z., Lu, Y., Pan, S., & Liu, Y. (2021). Efficient speech emotion recognition using multi-scale cnn and attention. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://doi.org/10.1109/icassp39728.2021.9414286>
- [21] Hou, Y., Kang, B., & Botteldooren, D. (2022). Audio-visual scene classification via contrastive event-object alignment and semantic-based fusion. 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP). <https://doi.org/10.1109/mmmsp55362.2022.9949447>

# The Influence of the Trademark on the Visual Identity of Social Networks Through User Experience

Nikola Jerković<sup>1</sup>, Ivana Berković<sup>1</sup>, Višnja Ognjenović<sup>1</sup>

<sup>1</sup> Technical Faculty "Mihajlo Pupin" Djure Djakovica bb, Zrenjanin, Republic of Serbia

nikolajerkovic13@gmail.com; ivana.berkovic62@gmail.com; visnjaognjenovic@gmail.com

## Abstract:

Social networks undeniably have a huge role in users daily lives. In such circumstances, understanding the impact of social media trademarks' design on users perception is invaluable. Consequently, this paper focuses on analyzing the subjective impression left on users by various trademarks. The obtained results provide a detailed insight into the significance of trademarks in shaping the visual identity of social networks, highlighting key factors that influence recognizability, aesthetic appeal, and emotional connection with users. Additionally, the paper offers guidelines for optimizing design to enhance user experience and improve the positioning of social networks in the digital space.

## Keywords:

social networks, user, design, trademark, subjective impression, visual identity

## 1. Introduction

Since humans became aware of their individuality, there has been a need to mark personal ownership and creations. Even in the early stages of civilization, people used various signs to preserve and express their identity. In line with this, the desire to highlight one's contribution was present in many fields, from masonry to literature and art. This trend has persisted through the ages and is evident today in simple examples, such as children signing their drawings, emphasizing their creative identity and authorship [1].

The concept of identity is associated with characteristics that indicate recognizability and uniqueness. Thanks to identity, it is possible to identify and differentiate, whether speaking of phenomena, people, organizations, or anything else. Visual identity, therefore, represents a set of graphic elements that enable recognizability. Such established recognizability in the context of a company refers to its public presentation and includes the impression it leaves on employees, clients, and partners. Generally, visual identity expresses the company's characteristics along with its values, core business idea, and ambitions, so it is important that it is both noticeable and consistent [1,2,3,4].

Companies have different requirements for defining their own visual identity, depending on their specific needs [5]. In general, the company's visual identity is determined by artistic elements that represent its visual constants, which are [5,6]:

- Trademark;
- Logotype;
- Company name;
- Typography;
- Colors;
- Layout system.

The trademark is the primary element in building and defining the company's visual identity and creating its recognizability [3]. In everyday usage, the term logo is often used as a synonym due to its simplicity and appeal, although in design contexts, it refers to the trademark [7].

The trademark, as a basic element of a company's visual identity, should be simple, attractive, and appealing to the observer [8], but also easily executable in various techniques, technologies, and sizes [3]. Its design is very important as it should represent and convey the appropriate idea related to the

company, which is why it is often highly abstract and symbolic. As the trademark is a universal language of recognizable symbols, a visual experience that speaks more than words, it can contain various segments, such as drawings, and even numbers and letters. Accordingly, with technological advancements, modern designers have a wide range of options for creating trademarks. There are various types of trademarks, with the most common being monograms, letters and numbers, geometric shapes and figures, illustrations, or pastiches as artistic works that mimic another artist's technique or style from another artistic period [8].

## 2. Research description

This research addresses the problem of building a recognizable visual identity for a social network. The specific subject of the study is the influence of the trademark as an element of visual identity on user perception. The goal of the research was to determine and compare the impression of social media trademarks based on the responses of participants of different ages and genders. The participants were tasked with expressing their opinions through a survey, rating the social media trademarks on a scale from 1 to 5, where a rating of 1 represents very poor, 2 insufficient, 3 neutral, 4 good, and 5 excellent. The trademarks of the following social networks were evaluated: Facebook, LinkedIn, Tumblr, Twitch, Twitter (X), Flickr, Badoo, Truth Social, Patreon, YuBo, YouTube, Instagram, TikTok, WhatsApp, Viber, Pinterest, Snapchat, Discord, Reddit, Threads.



Figure 1: Trademarks of social networks from research

The obtained research results are presented cumulatively through the display in Figure 2, which shows the calculated average ratings of individual trademarks according to all categories of respondents, taking into account the overall results, responses from individual gender and age groups, as well as their combinations.

For clarity, colors have been used to categorize the ratings as follows:

- Purple: rating  $\geq 4$ ;
- Blue:  $3 \leq \text{rating} < 4$ ;
- Red: rating  $< 3$ .

Additionally, colors were used to indicate the average ratings of trademarks that received the highest scores in any of the categories:

- Gold: trademark that received the highest rating in a specific category;
- Silver: trademark that received the second-highest rating in a specific category;
- Bronze: trademark that received the third-highest rating in a specific category.

In situations where the average ratings of trademarks coincide, all trademarks with the same average rating share the position.

Trademark	All respondents	Women	Men	Up to 25	From 25 to 50	Over 50	Women up to 25	Men up to 25	Women from 25 to 50	Men from 25 to 50	Women over 50	Men over 50
Facebook	3.4875	3.582375479	3.309352518	3.388888889	3.477272727	3.650943396	3.395061728	3.382716049	3.59375	3.166666667	3.75	3.272727273
LinkedIn	3.215	3.222222222	3.201438849	3.049382716	3.242424242	3.433962264	2.868888889	3.209876543	3.239583333	3.25	3.523809524	3.090909091
Tumblr	2.7125	2.827566207	2.496402878	2.728395962	2.636363636	2.783018868	2.962962963	2.49382716	2.708333333	2.444444444	2.833333333	2.590909091
Twitch	3.0225	2.911877395	3.230215527	3.154320968	2.840909091	3.047169811	2.888888889	3.419753066	2.802083333	2.944444444	3.05952381	3
Twitter (X)	2.8225	2.823754789	2.820143885	2.833333333	2.787878788	2.849056604	2.802469136	2.864197531	2.84375	2.638888889	2.821428571	2.954545455
Flickr	2.7625	2.892720307	2.517985612	2.407407407	2.825757576	3.226415084	2.283950817	2.530864196	3.020833333	2.305555556	3.333333333	2.818181818
Badoo	2.645	3.019157088	2.517985612	2.462962963	2.984848485	3.254716981	2.444444444	2.481481481	3.166666667	2.5	3.404761905	2.681818182
Truth Social	3.145	3.229850507	2.985611511	3.074074074	3.053030303	3.367924528	3.024691358	3.12345679	3.177083333	2.722222222	3.488095238	2.909090909
Patreon	2.7575	2.842911877	2.597122302	2.641975309	2.75	2.943396226	2.716049383	2.567901235	2.78125	2.666666667	3.035714286	2.590909091
YuBo	3.005	3.045977011	2.928057554	2.932098765	3.068181818	3.037735849	2.975308642	2.888888889	3.041666667	3.138888889	3.119047619	2.727272727
YouTube	4.2	4.187739464	4.230215883	4.209876543	4.196969697	4.188679245	4.190403827	4.258259259	4.177083333	4.25	4.226190476	4.045454545
Instagram	4.19	4.283524904	4.014388489	4.209876543	4.242424242	4.094339623	4.222222222	4.197530864	4.333333333	4	4.285714286	3.363636364
TikTok	3.6625	3.681992337	3.625899281	3.614814815	3.431818182	3.716981132	3.827160494	3.802469136	3.447916667	3.388888889	3.80952381	3.363636364
WhatsApp	4.0625	4.122605364	3.949640288	4.154320968	3.931818182	4.054905066	4.160493827	4.148148148	4.03125	3.666666667	4.19047619	3.681818182
Viber	3.995	4.049808429	3.892086331	3.864197531	4.083333333	4.054905066	3.913580247	3.814814815	4.083333333	4.083333333	4.142857143	3.863636364
Pinterest	3.7975	4	3.417266187	3.895061728	3.689393939	3.783018868	4.209876543	3.580246914	3.885416667	3.166666667	3.928571429	3.227272727
Snapchat	3.3	3.314176245	3.273381295	3.672839506	3.060606061	3.028301887	3.740740741	3.604938272	3.125	2.888888889	3.119047619	2.681818182
Discord	3.3425	3.180076628	3.647482014	3.790123457	3.083333333	2.981132075	3.50617284	4.074074074	3.03125	3.222222222	3.035714286	2.727272727
Reddit	3.485	3.45210728	3.54676259	3.574074074	3.378787879	3.481132075	3.320987654	3.827160494	3.427083333	3.25	3.607142857	3
Threads	3.085	3.206896552	2.856115106	3.074074074	3.090909091	3.094339623	3.172839506	2.975308642	3.270833333	2.611111111	3.166666667	2.618181818

Figure 2: The average ratings of trademarks at all levels of research

## 2.1. Discussion of research results

The provided overview allows for the analysis of the research results at all levels, including the participants' opinions on individual trademarks.

- Facebook's trademark received fairly consistent ratings, with slightly higher scores among the female audience compared to the male audience. Although there were no significant fluctuations in ratings across categories, Facebook's trademark did not stand out as much as some other trademarks.
- LinkedIn's trademark left a fairly average impression on the participants. It was rated highest among women over 50 years old, while it scored the lowest in the category of women under 25, where its average rating was slightly below 3.
- Tumblr's trademark did not meet the subjective impression of the participants, as its average rating in each category was below 3, which is unsatisfactory.
- Twitch's trademark generally oscillated around a rating of 3 across categories, sometimes slightly above and sometimes below this average. It was rated best among younger men under 25 years of age.
- Twitter's (X) trademark was rated quite evenly across categories, but since the average scores were below 3, it was perceived as insufficiently good by the participants.
- Flickr's trademark, according to the majority of participants, was deemed insufficiently good. It left a slightly better impression on the age group over 50 years old, specifically women over 50, as well as women aged 25 to 50, where its average rating was slightly above 3.
- Badoo's trademark received similar ratings to Flickr's, with the exception that it recorded an average score above 3 among female participants.
- Truth Social's trademark was generally rated solidly, around a 3, without leaving a strong impression on the participants. It received lower average ratings among male participants, particularly in the categories of men aged 25 to 50 and men over 50.
- Patreon's trademark did not meet the participants' expectations, as its average rating in all categories was below 3, except for the category of women over 50, where it received an average rating slightly above 3.
- YuBo's trademark did not show significant differences in average ratings across categories, though in some categories it exceeded 3, while in others it fell slightly below this average.
- YouTube's trademark was the only in this study to receive an average score above 4 across all categories. It was also the only platform whose trademark was ranked among the top 3 in every category, indicating that this trademark was particularly liked by the participants.
- Instagram's trademark closely followed YouTube's performance, also receiving an average score above 4 in almost all categories and ranking among the top 3. The only exception was the average rating from men over 50, who gave it a significantly lower score. Nevertheless, this trademark left a very positive subjective impression on the participants.
- TikTok's trademark was generally well-received by participants, with average ratings across all categories ranging between 3 and 4, converging around 3.5.

- WhatsApp's trademark was rated with outstanding average scores above 4 in most categories, with some fluctuations, and frequently ranked third in participants' opinions within specific categories. This clearly indicates that the trademark left a very positive impression on the participants.
- Viber's trademark had an average rating around 4, with slight fluctuations. In some categories, Viber's trademark was ranked among the top 3 based on participants' subjective impressions, indicating that it was well-received.
- Pinterest's trademark showed greater fluctuations in average ratings across participant categories. Although it did not score below 3 in any category, it received significantly lower ratings in some, compared to categories like female participants, where it scored 4, or female participants under 25, where it ranked right behind YouTube's trademark, as well as the category of participants under 25, where it ranked third in average score.
- Snapchat's trademark was rated as solid in most categories, except for the categories of older men aged 25 to 50 and over 50.
- Discord's trademark received quite different average ratings depending on the category, showing that it left very diverse impressions on participants. For instance, in the category of men under 25, its average rating exceeded 4, while in the generalized category of participants over 50 and the specific category of men over 50, its average rating was below 3.
- Reddit's trademark left a decent but not particularly strong impression on participants across all categories. Its average ratings ranged from 3 to 4, mostly falling within the lower half of that range.
- Threads' trademark, as a relatively new social network, was not particularly appreciated by participants, with its average ratings predominantly around 3, with slight deviations.

## 2.2. Conclusions based on research results

After analyzing the impressions of the trademarks, the following conclusions can be made:

- The most successful trademark belongs to the social network YouTube, followed by the trademarks of Instagram, WhatsApp, Viber, and then Pinterest.
- YouTube has the leading trademark according to male respondents at all levels of the research (including the category of all male respondents, the category of male respondents under 25, the category of male respondents aged 25 to 50, and the category of male respondents over 50), while this holds true for Instagram's trademark in the case of all the aforementioned categories of female respondents.
- In the category of respondents under 25, after the trademarks of YouTube and Instagram, the WhatsApp trademark received the highest ratings, with Pinterest also standing out, as it ranked second in the group of women under 25 based on its average trademark rating.
- On the other hand, in the category of respondents aged 25 to 50, alongside YouTube's and Instagram's trademarks, Viber's trademark was among the top three highest-rated, including within the separate categories of women aged 25 to 50 and men aged 25 to 50.
- The age group of respondents over 50 also expressed that YouTube's trademark is particularly appealing to them, praising Instagram's trademark as well, which was even ranked first among women over 50. However, men in this age group rated it much lower, so it did not make it into the top three highest-rated trademarks. On the other hand, respondents over 50 gave very consistent ratings to the trademarks of WhatsApp and Viber, which were also among the highest-rated. Women in this group favored the WhatsApp trademark, while men preferred the Viber trademark.

## 3. Recommendations for designing social media trademarks

Based on the presented research results, it is possible to provide general recommendations that can be applied in most cases, as they take into account the most common preferences based on the opinions of the participants in this study.

The initial recommendation is for the social media trademark to be visually simple. Considering that the new technological age has brought information overload, users no longer have the time or patience



for excessive detail, leading to an increasing trend of minimalism in trademark design. By removing unnecessary elements and retaining only what is essential, a more memorable visual identity is created. Limiting the color palette, lines, and shapes in the design of a trademark helps create a quick connection with the user, which is especially important in the context of social media. Additionally, simple and clear trademarks are very appealing as they evoke a sense of calmness and security in the observer.

The next important point is the coherence of the trademark design with the goals of the social network. The trademark should reflect the primary purpose of the network, so users immediately understand what it offers and form a connection with it. It is also crucial to creatively and intelligently combine elements that best represent the platform's essence to distinguish it clearly from others. Often, one effective graphic element that directly conveys the message to users is enough, and it should be arranged in a visually appealing way.

Color selection is also significant in the design of social media trademarks. Colors influence the emotional reactions of users, shaping their subjective perceptions of the network. From experience with designing social media trademarks, cheerful and energetic colors like red, pink, yellow, and orange, when combined with other synchronized elements, have produced excellent results, making these hues a strong recommendation as the foundation of a social network's visual identity. On the other hand, blue is the most commonly used color in social media design and is generally the most prevalent color in technology. However, for this reason, special attention should be paid to selecting a shade that sets the design apart from others. Unlike blue, its derivatives, purple and green, are less commonly used but are well-received because they help the social network's trademark stand out with inspiring and distinctive designs. It is also recommended to avoid excessively dark colors, as they often create a negative impression, especially in the context of social networks, which are primarily used for entertainment and relaxation. To enhance the impact of the design, it's not enough to choose a nice color; it must be aligned with the network's purpose and target audience. For platforms dominated by younger users, lighter and more vibrant colors are a good choice as they evoke excitement and energy. In contrast, for networks designed for professional purposes, whose users are mostly older, it is more appropriate to use neutral colors that convey stability and seriousness. Additionally, if the network has a significantly larger proportion of users from one gender, the colors used in the trademark design should be tailored to gender preferences. Female users tend to prefer softer shades, while males are generally drawn to bolder and clearer tones. Ultimately, while it is important to follow certain guidelines and trends, the design should not overly rely on generic colors commonly used by competitors. An innovative and thoughtful approach to color selection can significantly enhance the originality and recognizability of the trademark.

The shape of the trademark design should also be unique and aligned with the characteristics of the network it represents. Sometimes, shape alone can convey the essence of a social network and express it visually in a simple way. In practice, simple shapes have proven to be more memorable, so their combinations, alterations, or independent use should be carefully considered when creating a distinctive social media trademark. Moreover, forming an unusual shape that stands out from the rest makes the trademark memorable and noticeable.

Lines should be clean and simple to ensure the social media trademark looks clear and professional. In building a powerful visual identity, it is important to smartly choose the properties of the lines, such as thickness and style, so that they highlight the design and support its functionality. These should be combined in a way that avoids visual overload. Using very simple lines can create contrast, emphasize certain parts of the design, improve visual hierarchy, and direct users' attention, making it crucial to maintain precision, consistency, and uniformity in the use of lines in the trademark design.

Furthermore, when combining graphic elements with textual ones, it is advisable to use legible, simple, and consistent typography. Fonts in trademark design should be clear and easily recognizable, so it is recommended to use fonts with simple characteristics that allow for quick identification and readability. Limiting the number of fonts contributes to maintaining the professional look and recognizability of the trademark, so ideally, one font or a few related fonts should be used to avoid cluttering the design. Fonts not only convey visual information but also an emotional tone, so the chosen font must reflect the desired character and message of the social network. Rounded-edge fonts are often suitable for social media because they appear approachable and friendly, while sharp-edge fonts are better for platforms that emphasize seriousness and sophistication.

Although these recommendations provide basic guidelines for designing social media trademarks, it is essential to remember that every trademark is specific and unique, requiring a tailored approach. The final design should always be adapted to the specific message and goals of the social network, as well as the desires, needs, and expectations of the users. Therefore, it is necessary to balance following these guidelines with creatively adapting to each case to ensure the success of the trademark.

## 4. Conclusion

With the increasing number of users spending time on social media, their visual identity is becoming ever more important. A unique trademark significantly contributes to highlighting the social network, creating an opportunity to connect with users, build loyalty, and achieve business goals, as clearly demonstrated by this research. Additionally, based on the average ratings of social media trademarks from respondents, it can be concluded that there is considerable room for improvement and enhancement of impressions, as well as the development of a more effective and powerful visual identity. This can also serve as a suggestion for a more detailed consideration and careful selection of trademark designs. Popular networks generally received excellent ratings from respondents for their trademarks, but this does not mean they should strictly adhere to the current design in the future. Designers, in general, should not rigidly follow guidelines in creating trademarks but should adapt to the specific needs of the platform and its users. Design is an art, and as such, it knows no boundaries or rules. What must always be kept in mind, however, is the fact that subjective impressions vary greatly, as confirmed by this research, meaning there is no established formula for success when it comes to individual experiences. Based on past experiences, it is only possible to observe certain aesthetic and emotional norms that apply on a broader scale, making it advisable to follow them while avoiding elements that are unappealing to the majority of users. In any case, only through continuous trend monitoring, testing of innovative and creative ideas, and carefully listening to user feedback can the full potential of a social network's visual identity be harnessed, offering the chance to achieve and maintain success in the digital market.

### References:

- [1] M. Juričić, Vizuelni identitet kao element uspješnog upravljanja i poslovanja tvrtke, Undergraduate thesis, Sveučilište Josipa Jurja Strossmayera u Osijeku, Filozofski fakultet, Osijek, 2015.
- [2] Ž. Biškup, Vizualni identitet fotografskog studija Oblak, Undergraduate thesis, Sveučilište Sjever, Varaždin, 2018.
- [3] A. Mikić, Vizuelni identitet organizacije u funkciji komunikacije, in: 2rd International Scientific Conference on "POWER OF COMMUNICATION 2013", Banja Luka : Panevropski univerzitet Apeiron, 2013, pp. 187-199. [https://www.researchgate.net/profile/Antoaneta-Vassileva/publication/334612477\\_Paradoxes\\_of\\_the\\_Global\\_Marketing\\_Communication\\_Peculiarities\\_Improbabilities\\_Opposing\\_Attitudes\\_Conference\\_Proceedings\\_Power\\_of\\_Communication\\_2013\\_Belgrade/links/5fe7969b45851553a0f5b16a/Paradoxes-of-the-Global-Marketing-Communication-Peculiarities-Improbabilities-Opposing-Attitudes-Conference-Proceedings-Power-of-Communication-2013-Belgrade.pdf#page=189](https://www.researchgate.net/profile/Antoaneta-Vassileva/publication/334612477_Paradoxes_of_the_Global_Marketing_Communication_Peculiarities_Improbabilities_Opposing_Attitudes_Conference_Proceedings_Power_of_Communication_2013_Belgrade/links/5fe7969b45851553a0f5b16a/Paradoxes-of-the-Global-Marketing-Communication-Peculiarities-Improbabilities-Opposing-Attitudes-Conference-Proceedings-Power-of-Communication-2013-Belgrade.pdf#page=189)
- [4] K. Horvat, Dizajn i primjena vizualnog identiteta studija Millstone, Undergraduate thesis, Sveučilište Sjever, Varaždin, 2023.
- [5] D. Marković, D. Cvetković, Osnovi grafičkog dizajna, Univerzitet Singidunum, Fakultet za informatiku i menadžment, Beograd, 2009.
- [6] F. Arnerić, Web aplikacija za izradu vizuelnog identiteta kompanije, Bachelor's thesis, Univerzitet Singidunum, Departman za informatiku i računarstvo, Beograd, 2011.
- [7] V. Bižić, F. Kuduz, Značaj logoa kao mogućnosti dobrog brendiranja na primeru Kyokushinkai karate simbola, in: 2rd International Scientific Conference on "POWER OF COMMUNICATION 2013", Banja Luka : Panevropski univerzitet Apeiron, 2013, pp. 19-25. [https://apeiron-uni.eu/wp-content/uploads/2020/04/MOC-KOMUNIKACIJE-Zbornik-radova\\_2013.pdf#page=21](https://apeiron-uni.eu/wp-content/uploads/2020/04/MOC-KOMUNIKACIJE-Zbornik-radova_2013.pdf#page=21)
- [8] T. Ogrinec, Redizajn vizualnog identiteta branda Čokolešnik, Undergraduate thesis, Sveučilište Sjever, Varaždin, 2017.

# The Inheritance Relationship in the Teaching of Object-Oriented Programming

Aleksandar Kupusinac

*University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia*

*sasak@uns.ac.rs;*

## Abstract:

The inheritance represents one of the most important concepts of object-oriented programming, which allows obtaining a new class based on the content of the old class (parent class). Based on that, inheritance can be applied as a powerful tool that provide multiple use of programming code. In this paper, we will discuss few important differences that exists between implementations of inheritance in C++, Java and C# programming languages.

## Keywords:

C++, Java, C#, inheritance, teaching

## 1. Introduction

The first topic in object-oriented programming teaching is – class and object. Class is a model of class concept, while object is a model of individual concept [1]. The content of the class consists of fields and methods, where the field represents the model of descriptive and the method the model of operational property. An object is an instance of a class, which has identity, state and behavior. The object has a unique identity, i.e. the name uniquely identifies that object in an object-oriented program. The state of the object is determined by the current field values. Methods can change the state of an object. The life of an object is one path through the state space. Object-oriented program is a model of real-world environment where object “lives”, changes states and behaves according to class methods.

The second topic in object-oriented programming teaching is – class relationships. There are complex relations between class concepts in the real-world environment. Class relationships represent a model of this relations, e.g. relation between class subjects *<Teacher>* and *<Subject>* could be described by sentences: *<Teacher> teaches <Subject>* and *<Subject> is taught by <Teacher>*, or e.g. relation between class subjects *<Car>* and *<Motor>* could be described by sentence: *<Car> contains <Motor>*, or e.g. relation between class subjects *<Person>* and *<Student>* could be described by sentence: *<Student> is a <Person> that study at a university*.

There are numerous types of class relationship, that can be divided into two groups [2]:

- Client relationships,
- Inheritance.

The client relationship describes situation where one class (client) uses services of another class (server). The most used client relationships are association, aggregation, and composition. Association describes a two-way relationship between two classes, where both classes can participate as client or server, e.g. *<Teacher> teaches <Subject>* and *<Subject> is taught by <Teacher>*. Owner-component relations are modeled by aggregation or composition, where one class (owner) contains another (component). Aggregation implies that the lifetime of the owner and component are independent, e.g. *<Car> contains <Motor>*. On the other hand, composition implies that the owner and component are simultaneously constructed and destructed, *<Cylinder> contains <Base> and <Wrapper>*.

The inheritance implies creation of new class as an extension of parent class, where the new class presents a special kind of parent class, e.g. *<Student> is a <Person> that study at a university*, where *<Student>* is a new class derived from parent class *<Person>*. The content of the parent class can be downloaded without or with modification, with the possibility of adding new content. This paper analyzes few important differences that exists between implementations of inheritance in C++, Java and C# programming languages.

## 2. Syntax

There are different implementations of inheritance in numerous programming languages. In C++, the derivation of class B from class A is written as follows:

```
class A {
 // fields
 // methods
};

class B : <access-specifier> A {
 // fields
 // methods
};
```

The word <access-specifier> can be private, protected or public. When <access-specifier> is private, then all inherited content become private. In case when <access-specifier> is protected, then all inherited content become protected, except content that is already private where access right will stay private. If <access-specifier> is public, then access rights to all inherited content will not change. Field in the derived class can be inherited or added, while method can be inherited, overridden or added. By default, method is not virtual, except when it has word `virtual` at the beginning.

Programming language Java uses word `extends` instead `:` (colon) [4]. There is no <access-specifier>, since each inheritance is public by default. In Java, by default, each method is virtual. When method is modified, then in the derived class is used word `overridden` at the beginning. The derivation of class B from class A is written as follows:

```
class A {
 // fields
 // methods
}

class B extends A {
 // fields
 // methods
}
```

Programming language C# uses `:` (colon) instead word `extends` [5]. In C#, by default method is not virtual and word `virtual` at the beginning, like in C++. When method is modified, then in the derived class is used word `overridden` at the beginning. Each inheritance is public by default. The derivation of class B from class A is written as follows:

```
class A
{
 // fields
 // methods
}

class B : A
{
 // fields
 // methods
}
```

### 3. Private inheritance

The programming language C++ provides the possibility of realizing private inheritance, which means that all inherited content will become private. There are situations where private inheritance is a very useful tool, such as the deriving class `Stack` from class `List`:

```
class List {
 // fields
 // methods add(i,e1), read(i), remove(i)
};

class Stack : private List {
 // methods
 // push() {add(1,e1)}
 // pop() {remove(1)}
 // top() {read(1)}
};
```

On the other side, private inheritance allows class deriving that do not make sense, i.e. where any class can be derive from any class, e.g. class `Student` from class `Triangle`.

Syntax of programming languages Java and C# does not contain `<access-specifier>` for inheritance and assumes that each inheritance is public.

### 4. Multiple inheritance

Programming language C++ contains multiple inheritance, where class can be derived from two or more base classes. Eg. class `C` is derived from classes `A` and `B`:

```
class A {
 // fields
 // methods
};

class B {
 // fields
 // methods
};

class C : public A, public B {
 // fields
 // methods
};
```

The need to use multiple inheritance exists when one class concept has the functionalities of two or more concepts. Eg. class concept `<AlarmClock>` has functionalities of class concept `<Alarm>` and class concept `<Clock>`. On the other hand, multiple inheritance can produce few problems. Let us consider the first problematic situation with multiple inheritance. Eg. if classes `A` and `B` both contain the same field `x`, then class `C` derived from classes `A` and `B` will have two fields `x`:

```
class A {
 // field x
 // methods
};
```

```
class B {
 // field x
 // methods
};
```

Now, let us consider the next problematic situation with multiple inheritance. Eg. if class A is a mutual parent of classes B and C, then class D derived from classes B and C will have two parent parts which originate from the same class A:

```
class A {

};

class B : public A {

};

class C : public A {

};

class D : public B, public C {

};
```

Syntax of programming languages Java and C# does not contain multiple inheritance.

## 5. Conclusions

Inheritance is a very important topic in teaching object-oriented programming. There are numerous challenges that must be overcome to adequately present inheritance to students. By that, this paper analyzed few important differences that exist between implementations of inheritance in C++, Java and C# programming languages.

### Acknowledgment:

This work was supported by the Faculty of Technical Sciences in Novi Sad as part of the implementation of the project entitled "Improvement of teaching processes, scientific and artistic research with challenges followed by new models of funding and ranking of scientific-research organizations".

### References:

- [1] A. Kupusinac, *Programming Language C++, (in Serbian)*. Novi Sad: FTN Tehnička edicija, 2020.
- [2] B. Meyer, *Object-Oriented Software Construction, (second edition)*. Prentice Hall, 1997.
- [3] D. Malbaški, *Object-oriented programming through programming language C++, (in Serbian)*. Novi Sad: FTN Tehnička edicija, 2020.
- [4] B. Eckel, *Thinking in Java, (4<sup>th</sup> edition)*. Prentice Hall PTR, Upper Saddle River, NJ, United States, 2005, ISBN: 9780131872486.
- [5] L. O'Brien and B. Eckel, *Thinking in C#*. Prentice Hall PTR, Upper Saddle River, NJ, United States, 2003, ISBN 9780130385727.

# Testing of the e-security systems in the automotive industry

Predrag Novokmet<sup>1</sup>, Vesna Makitan<sup>1</sup>, Dragana Glušac<sup>1</sup>, Eleonora Brtka<sup>1</sup>, and Nikola Jovanov<sup>1</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin" Djure Djakovica bb, Zrenjanin, Republic of Serbia

*predrag.novokmet@tfzr.rs; vesna.makitan@tfzr.rs; dragana.glusac@tfzr.rs; eleonora.brtka@tfzr.rs; nikola.jovanov@tfzr.rs*

## Abstract:

In the modern automotive industry, e-security technologies play a crucial role in enhancing vehicle and passenger safety. This paper examines various technologies and applications of these technologies that contribute to automotive safety. The main focus is on systems such as Electronic Stability Control (ESC), Adaptive Cruise Control (ACC), autonomous driving systems, virtual cockpit, keyless entry, and others. This paper deals with the advantages and challenges of implementing these systems, and their impact on reducing traffic accidents.

The aim of the research was to establish the extent to which these technologies can contribute to increasing the safety of vehicles and their passengers, i.e. to the greater safety in traffic. A detailed analysis and comparison of these systems can help engineers and car manufacturers to improve existing technologies and develop new ones, even more efficient and thus to contribute to the development of the entire e-business. Therefore, as expected research results were guidelines for further development and implementation of advanced e-security technologies. The paper concludes that these technologies have significantly contributed to increasing road safety, reducing the number of injuries and fatalities, and providing economic benefits to the industry and consumers.

## Keywords:

E-Security, Automotive Industry, Vehicle Safety, IT Technologies

## 1. Introduction

The automotive industry is constantly developing, and technology plays a key role in improving vehicle and passenger safety. The integration of e-security systems has become necessary in order to reduce the number of accidents and increase traffic safety.

E-security in the automotive industry includes a set of technologies and systems designed to improve the safety of vehicles and their passengers. These systems use advanced electronic components and software algorithms to recognize and respond to potentially dangerous traffic situations, reducing the risk of accidents and adverse consequences [1, 2].

E-security includes various systems that work together to provide comprehensive protection. Electronic Stability Control (Electronic Stability Program or Electronic Stability Control – ESC) is one of the key e-security systems, which uses sensors to monitor vehicle movement and automatically adjusts braking on individual wheels to prevent skidding. Adaptive Cruise Control (ACC) uses radar or laser sensors to maintain a safe distance between vehicles, automatically adjusting the driving speed. [3]

Autonomous driving systems are the most advanced form of e-security, using a combination of cameras, radars, lidars and advanced algorithms to recognize the environment and control the vehicle independently without driver intervention. These systems can recognize obstacles, pedestrians and other road users, and react in time to avoid accidents.

The instrument panel, i.e. the virtual cockpit, is also part of the e-security, providing the driver all the key information in an easy-to-view manner. These systems use digital displays to display speed, navigation, vehicle status information and warnings about potential hazards, reducing the risk of drivers' error.

The Keyless Entry system allows drivers to unlock and start the vehicle without the need for a physical key, using smart keys or fobs that communicate with the vehicle via radio signals. This system contributes to practicality as well as safety.

Together, all these technologies present an e-security system, which aims to make driving safer and more efficient. Through the integration of this advanced system, the automotive industry strives to reduce the number of accidents, injuries and deaths on the roads, providing drivers and passengers with a greater level of safety. [4]

With an increasing number of vehicles on the roads, the application of advanced e-security technologies becomes necessary to ensure the safety of all road users.

The mentioned technologies: electronic stability control (ESC), adaptive cruise control (ACC), autonomous driving systems, instrument panels and Keyless Entry systems are the basis of modern vehicles. [5]

The idea behind the research in this paper arose from everyday use and the observation that some e-security systems in cars do not contribute in the way they should. Although they increase security, they often cause new problems that can be more serious than the ones they solve.

This paper will present the basic characteristics of the mentioned e-security technologies, as well as their advantages and challenges in implementation. Also, with this research it will be established the extent to which these technologies can contribute to increasing the safety of vehicles and their passengers, i.e. greater safety in traffic. A further detailed analysis and comparison of these systems can help engineers and car manufacturers to improve existing technologies and develop new ones, even more efficient. Therefore, as expected research results, guidelines for further development and implementation of advanced e-security technologies will be presented.

In this way, the presented research will contribute to a better understanding and implementation of e-security systems, with an insight into their impact on reducing the number of traffic accidents and increasing overall road safety.

## **2. Theoretical background – Historical overview of security systems in the automotive industry**

The first cars were equipped with the very basic mechanical security systems. In the early stages of the automotive industry development, the emphasis was on enabling the basic functionalities of the vehicle more than on the passengers safety. As the industry developed, the need for safety measures improvement in order to reduce the risk of injuries in the traffic accidents became clear. One of the first significant steps in the field of safety was the introduction of seat belts. The seat belt, which first appeared in the 1950s, has become a standard part of the equipment of most cars. It has drastically reduced the number of injuries and deaths by keeping passengers in their seats during crashes, preventing them from being ejected from the vehicle. After that, safety systems in cars underwent significant development, moving from basic mechanical solutions to sophisticated electronic systems. The 20th century was marked by numerous innovations that improved the overall safety of vehicles [6].

Another key innovation was the introduction of airbags. Airbags were first installed in cars in the 1970s and became standard in the 1980s. These systems are designed to inflate automatically in the event of a crash, providing additional protection to occupants and reducing the risk of head and chest injuries [7].

One of the first major innovations was the introduction of the Antilock Braking System (ABS) during the 1970s and 1980s. The ABS is designed to prevent the wheels from locking during sudden braking, allowing the driver to maintain control of the vehicle. This system has significantly reduced the number of accidents caused by loss of control during braking, especially on slippery or wet roads [8].

The next step in the safety systems evolution was the development of electronic stabilization systems (Electronic Stability Control – ESC). This system was introduced in the late 1990s and early 2000s. The ESC system uses sensors to monitor the vehicle's movement and, in the event of control loss, automatically applies brakes to individual wheels to stabilize the vehicle. The ESC has become



crucial in preventing vehicle rollover and control loss, especially in sharp curves and sudden direction changes [9].

The development of these electronic systems represented a significant advance in the automotive industry, laying the foundations for today's advanced e-security technologies that use a combination of sensors, computers and communication networks to provide an even higher level of safety and protection for passengers [10].

The first Adaptive Cruise Control – ACC systems appeared in the mid-1990s, with initial research conducted by car manufacturers such as Mitsubishi and Toyota [6]. The ACC is an advanced driver assistance system that uses sensors and radars to monitor traffic in front of the vehicle and automatically adjusts the vehicle's speed to maintain a safe distance from the vehicle in the front [7, 11].

The autonomous driving system works through a combination of multiple technologies and sensors that allow the vehicle to perceive the environment, make decisions and perform actions instead of the driver. Early experiments from the 1980s (Carnegie Mellon University with the Navigation Laboratory project) contributed to the development of autonomous vehicles. Then in the 1990s, DARPA (Defense Advanced Research Projects Agency) launched a series of challenges known as the DARPA Grand Challenge, with the goal of developing autonomous vehicles for military purposes. Finally, in the first decade of the 21st century, major technology and automotive giants such as Google (now Waymo), Tesla, Uber and many others began to actively develop and test autonomous vehicles. Google first introduced its autonomous vehicle in 2009, which used a combination of lidar, radar, cameras and advanced algorithms to navigate without driver intervention [12].

The instrument panel (virtual cockpit) is one of the key parts of the car's interior, providing the driver with all the necessary information about the vehicle's operation and driving conditions. The first cars had very simple instrument panels, with only a few basic indicators, such as a speedometer and a fuel gauge [11]. During the 20th century, instrument panels became more and more complex, adding more gauges and indicators to give the driver better information about the vehicle's performance [13]. Modern systems include advanced features such as augmented reality (AR) that can display information directly on the windshield, increasing safety and driving comfort [7].

The Keyless Entry system is a technology that allows drivers to unlock and start their vehicle without a physical key. This technology significantly improves the comfort and safety of the vehicle, reducing the risk of theft and facilitating access to the vehicle [6]. The first systems date back to the 1980s and were relatively simple, allowing drivers to unlock the doors with a limited-range remote control. Modern keyless entry systems today are extremely secure and adaptable. They integrate smartphones allowing drivers to use mobile applications to access the vehicle and control various functions remotely [11].

### **3. Research methodology**

#### **3.1. Research goal and hypothesis**

The aim of this research is to determine whether e-security technologies in the automotive industry provide drivers with greater safety and improve traffic safety, as well as what drivers' expectations are from these technologies. It is expected to determine the level of driver awareness of the benefits of technologies such as Electronic Stability Control (ESC), Adaptive Cruise Control (ACC), Autonomous Driving Systems, digital instrument panel and Keyless Entry systems, i.e. that these technologies increase driving safety and comfort. Also, it is expected to determine drivers' willingness to use abovementioned technologies. From the primary goal comes the evaluation of benefits, i.e. assessment of whether modern versions of these systems have contributed to increasing vehicle safety and traffic accidents reduction.

The main hypothesis of this research focuses on technologies and their applications that have contributed to increased safety in the automotive industry and it states: "Modern technologies and their implementation can contribute to increasing safety in the automotive industry".

### 3.2. Research method and research sample

The method used in this research is surveying. The aim of the survey was to determine respondents' views on various e-security technologies in the automotive industry, with a special focus on their effectiveness, implementation and perception among professionals.

The survey was designed to provide a deeper insight into how experts in the automotive industry value these technologies, what advantages and disadvantages they recognize, and how their implementation affects the vehicles and passengers safety.

A targeted sample composed of experts from the automotive industry who work in large global corporations based in Serbia, such as ZF Serbia, Continental and other significant companies, was selected for the respondents. The reasons for choosing these respondents are multiple:

- **Years of experience:** The respondents have many years of experience in the automotive industry, which allows them to provide a deep understanding and insight into the evolution and implementation of e-security technologies.
- **Specific expertise:** As professionals working on the development and implementation of these technologies, they are most familiar with the technical details, challenges and benefits that these innovations bring.
- **Knowledge of the issues:** By working in large global corporations, the respondents are familiar with global trends and practices in the automotive industry, which allows them to provide relevant answers. In addition, their expertise allows insight into the actual situation and potential improvements in the application of new technologies.

The survey was an online anonymous survey in order to allow respondents to freely and honestly express their views without fear of negative consequences or pressure. The survey included a sample of 31 respondents, which in the case of this research is considered representative, because they are experts from the automotive industry in Serbia who deal with the application of e-security systems.

## 4. Research results and discussion

A summary of the survey results is given in table 1, where it may be see the mean value of all the ratings of the mentioned systems, their mutual relationship and importance for vehicle and driver safety. The values in the Total Score column are obtained as the sum of the product of the number of respondents and the score divided by the number of respondents.

**Table 1:**

Summary of survey results

Question No.	Score					E-security technology	Influence	Total Score
	1	2	3	4	5			
1.	6,9	6	67,7	9,7	9,7	Keyless entry	Theft protection	3,06
2.	0	0	6,5	9,7	83,9	ESC	Vehicle safety	4,77
3.	3,2	0	64,5	32,3	0	All	Theft protection	3,26
4.	3,3	3,6	6	74,2	12,9	Autonomous driving	Traffic accidents	3,87
5.	0	0	29	71	0	Autonomous driving	Traffic accidents	3,71
6.	3,2	3,2	87,1	6,5	0	Autonomous driving	Usability	2,97
7.	0	3,1	6,5	19,4	71	ESC	Vehicle control	4,58
8.	3,1	3,1	6,5	87,1	0,2	All	Vehicle safety	3,77
9.	16,1	74,2	6,5	3,2	0	Virtual cockpit	Driver safety	1,96
10.	6,5	6,4	6,5	77,4	3,2	Management assistance	Shopping assistance	3,61
11.	0	3	6,5	83,9	6,7	All	Mortality reduction	3,93
<b>Average</b>	3,84	9,33	26,66	43,13	17,05	-	-	3,58

Based on the collected data from the survey and the results analysis, it can be concluded that these technologies have generally significantly contributed to increasing safety, however, not all technologies have contributed to the same extent.

Research has shown that the hypothesis is not fully confirmed with all security systems. In particular, the autonomous driving systems, ESC and ACC confirmed the hypothesis, while with the Keyless Entry and Virtual Cockpit systems, the results showed that these technologies did not contribute at the same level of efficiency as the other systems. Respondents expressed concern about security risks associated with this system, such as vehicle theft through relay attacks [14]. More than half of respondents (67.7%) believe that Keyless Entry systems increase the risk of vehicle theft, while only 12.9% believe that this system, minimally or not at all, increases the risk of vehicle theft. This indicates the need to further improve the security features of these systems in order to increase user confidence. The problem arose because manufacturers left the possibility of easier illegal access to the vehicle, by using signal boosters [14].

Although Virtual Cockpit technology offers numerous advantages in terms of visibility and customization of information [15], respondents expressed doubts about its reliability and potential for driver distraction. The results show that 93% of the respondents believe that the Virtual Cockpit had a minimal positive impact on increasing safety, while only 7% of the respondents believe that the Virtual Cockpit had a moderate or significant positive impact on safety.

The results of the survey showed that the majority of respondents believe that ESC (Electronic Stability Control) contributes to vehicle safety to an extremely large extent. In particular, 83.9% of respondents believe that ESC significantly contributes to maintaining vehicle control, especially in bad weather conditions. Autonomous driving systems are also recognized as significant contributors to safety. The results of the survey show that 74.2% of respondents believe that autonomous driving systems will significantly reduce the number of traffic accidents in the future. This indicates a high degree of confidence in these technologies and their ability to improve overall road safety.

If the summary results in table 1 are observed, it can be concluded that the average rating of all e-security systems is generally high, because  $43.13 + 17.05 = 60.18\%$  of the respondents gave the highest ratings (4 and 5 respectively) to the tested e-security systems. Also, the Total Score column indicates that the best rated e-security system is the ESC, which is recognized as the system that users believe contributes the most to vehicle and driver safety. On the other hand, the lowest grade, i.e. Virtual cockpit is least trusted by users. The above also supports what was previously said, as far as the confirmation of the research hypothesis is concerned.

The introduction of advanced e-security technologies in the automotive industry is driven by the need to increase road safety, reduce the number of traffic accidents and their consequences, and improve the overall driving experience. Advanced technologies, such as Electronic Stability Control (ESC) and Adaptive Cruise Control (ACC), allow cars to intelligently react to changes in the environment and the behavior of other vehicles, thus significantly increasing safety [8, 10, 16]. The integration of Virtual Cockpit and Keyless Entry systems not only aim to increase comfort and practicality for drivers, but also aim to improve safety through better information and protection against theft [10, 17].

## 5. Conclusions

This paper examines the impact of modern technologies and e-security systems on the safety of drivers and vehicles in the automotive industry. The research focus was on some of the key e-security systems such as Electronic Stability Control (ESC), Adaptive Cruise Control (ACC), Autonomous Driving Systems, Virtual Cockpit and Keyless Entry.

In order to confirm the research hypothesis, research was conducted by surveying experts and professionals involved in the development and implementation of these systems in the automotive industry, enabling the collection of relevant data on their effectiveness and perception within the industry. The results were analyzed to determine whether these systems met the safety objectives set.

Future research could include a wider sample of respondents, specifically including end users of e-security technologies, in order to gain a more comprehensive insight into their actual system use and experience. Such an expanded sample would enable a comparison of the views of professionals and end users, which would provide valuable information for the further development and optimization of these technologies.

The research results confirmed that modern e-security systems significantly improve vehicle safety. Electronic Stability Control (ESC) showed high efficiency in preventing skidding and loss of vehicle control, while Adaptive Cruise Control (ACC) enabled maintaining a safe distance from other vehicles, reducing the risk of chain collisions. Autonomous driving systems allow vehicles to move independently and react to traffic situations, further increasing driving safety. The Virtual Cockpit provides drivers with intuitive and customizable information, but does not contribute to safety, and the situation is very similar with the Keyless Entry system, which improves comfort, but reduces vehicle safety and increases the theft risk.

Based on the above, some recommendations can be made for the further development of technologies that would contribute to increasing the safety of vehicles and drivers: improving the safety of the Keyless Entry system; further development of autonomous driving systems; driver education on the benefits and proper use of e-security technologies.

#### **Acknowledgment:**

This research is financially supported by Ministry of Education and Science of the Republic of Serbia under the project number TR32044 "The development of software tools for business process analysis and improvement", 2011-.

#### **References:**

- [1] Đukić, M., & Petrović, S. (2020). Tehnologije e-zaštite u automobilske industriji. Beograd: Tehnički fakultet.
- [2] Marković, D. (2021). Primena elektronskih sistema u automobilima. Kragujevac: Univerzitet u Kragujevcu.
- [3] Blanco, M., & Hankey, J. (2019). Handbook of Intelligent Vehicles. Springer.
- [4] Pérez, E., & Torres, R. (2018). Automotive Safety Systems. Berlin: Springer – Verlag.
- [5] Rogers, D. (2019). The Automated and Connected Vehicle: Status, Business, and Technical Challenges. London: Institution of Engineering and Technology.
- [6] Nikolić, S. (2017). Automobilska industrija i nove tehnologije. Beograd: Saobraćajni fakultet.
- [7] Pavlović, I. (2019). E-zaštita u automobilske industriji: Pregled i perspektive. Novi Sad: Univerzitet u Novom Sadu.
- [8] Petrović, V. (2020). Elektronski sistemi u vozilima: Bezbednosni aspekti. Beograd: Univerzitet u Beogradu.
- [9] Savić, M. (2018). E-zaštita u savremenim vozilima: Tehnologije i primene. Niš: Univerzitet u Nišu.
- [10] Bishop, R. (2020). Intelligent Vehicle Technology and Trends. Norwood, MA: Artech House.
- [11] Jovanović, B. (2018). Napredne tehnologije u savremenim automobilima. Novi Sad: Fakultet tehničkih nauka.
- [12] Gietelink, O., Ploeg, J., & De Schutter, B. (2018). Development of Advanced Driver Assistance Systems with Vehicle Hardware-in-the-Loop Simulations. IEEE Transactions on Intelligent Transportation Systems.
- [13] Kovačević, M. (2019). Bezbednost saobraćaja i e-zaštita. Zrenjanin: Tehnički fakultet "Mihajlo Pupin".
- [14] Miller, C., & Valasek, C. (2017). Car Hacking: A Practical Guide to Exploiting Software and Hardware. O'Reilly Media.
- [15] Rajamani, R. (2020). Vehicle Dynamics and Control. Boston: Springer.
- [16] Vučković, J. (2020). Adaptivni tempomat i sigurnosni sistemi u vozilima. Beograd: Saobraćajni fakultet.
- [17] Stojanović, D. (2021). Sistemi bezbednosti u automobilima: Elektronska kontrola stabilnosti. Novi Sad: Fakultet tehničkih nauka.

# Review of Different Types of Computer-Aided Engineering

Predrag Dašić<sup>1</sup>, Raul Turmanidze<sup>2</sup>, Vazha Shilakadze<sup>2</sup>, and Giorgi Popkhadze<sup>2</sup>

<sup>1</sup> *Engineering Academy of Serbia (IAS), 11000 Belgrade, Serbia*

<sup>2</sup> *Georgian Technical University (GTU), Faculty of Transportation and Mechanical Engineering, Tbilisi, Georgia*  
[inform@gtu.ge](mailto:inform@gtu.ge); [dasicp58@gmail.com](mailto:dasicp58@gmail.com); [popxadzegiorgi@gmail.com](mailto:popxadzegiorgi@gmail.com)

## Abstract:

Computer-aided engineering (CAE) is a term used to describe the various technologies and methods used for engineering analysis at all stages of development and manufacturing of products. CAE is widely used to simulate real-world conditions, optimize designs and validate product performance, with the aim to improve the overall engineering process. For these technologies and methods, numerous standards and specialized software systems and tools have been developed. Additionally, several different types of CAE technologies and various forms of CAE technology integration with other similar CAX technologies have been created. This paper provides an overview and analysis of different types of CAE technologies and the various forms of CAE technology integration with other similar CAX technologies.

## Keywords:

Computer-aided engineering (CAE), computer-aided design (CAD), computer-aided manufacturing (CAM), computer-aided quality (CAQ), computer-aided robotics (CAR).

## 1. Introduction

Activities related to modern design, simulation, testing, manufacturing, quality control, and logistics of new products are unimaginable without computer-aided technologies, or computer-aided everything (CAX), such as CAD, CAM, CAE, CAPP, CAQ, CAS, CIM, PLM, and so on. All types of CAX technologies employ a large number of techniques and methods, as well as software systems and tools, to solve engineering problems in various scientific fields and industries.

The application of different CAX variants that utilize a multitude of technologies, methods, software systems, and tools enables:

- Rapid achievement of a final product version based on simulation and testing of a virtual prototype (VP) or rapid prototyping (RP), which significantly reduces the time needed for simulation, production, and testing of product prototypes, thus accelerating the introduction of a new product version to the market;
- Efficient and fast production and quality control of the new product, significantly shortening the time required for manufacturing and quality control;
- Creation of a reliable product for practical and real-world operating conditions;
- Efficient planning and management of the manufacturing process and logistics and marketing activities related to transportation and market placement of the product;
- Significant advantages in design, planning, production, quality control, and logistics of the new product, among others.

All the above factors directly lead to a significant reduction in the time needed to bring a new product version to market and lower production costs while ensuring product quality and reliability in real-world operating conditions.

Within the activities of simulation and/or prototyping, simulation, testing, manufacturing, and quality control of the new product, engineering analysis is performed using various technologies and methods within CAE (Computer-Aided Engineering), which is considered one of the most important CAX technologies.

CAE uses software systems and tools to solve engineering problems across various scientific fields and branches of engineering, in diverse areas of medicine, and in the design and manufacturing of devices and equipment across different industries (mechanical, automotive, electrical, food, agriculture, etc.), employing various technologies, techniques, and methods of engineering analysis.

The aim of this paper is to present the basic types (variants) of CAE technologies, along with the application and advantages of integrated working frameworks, platforms, and CAE software systems combined with other similar CAx technologies.

## 2. Computer-Aided Technologies

Computer-aided technologies (computer-assisted technologies) or computer-aided everything (computer-assisted everything) or computer-aided "X" or abbr. CAx is a general term for technologies that use computer technologies (hardware and software) to aid in design, manufacture and exploitation of products [1-4]. CAx is a collective term for a large number of various computer-aided technologies (systems and processes), relating to the aspects of development, design, analysis, simulation, manufacturing planning, manufacturing, testing and quality assurance of products.

CAx technologies represent a set of hardware components consisting in computer and interactive devices and software systems and tools by means of which are most often automated activities during the development of a new product and preparation for its manufacturing (Figure 1) [4]. Hardware components and software systems and tools within CAx technologies are in direct interaction with data, knowledge and human's activities.

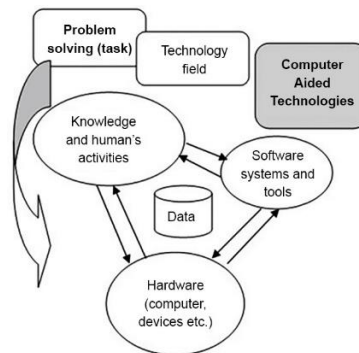


Figure 1: The main components of computer-aided technologies (CAx) [4]

Advanced CAx software systems and tools merge many different aspects of the product lifecycle management (PLM) [5], including design, analysis, production planning, manufacturing of products, etc. They can also be integrated with other computational systems for management and planning of production systems and processes, such as EDM (electronic document management), ERP (enterprise resource planning), MRP (material resource planning), MRP II (manufacturing resource planning), NPD (new product development), PDM (product data management), PDM (product definition management), PPC (production planning and control), RPD (rapid product development), SDM (simulation data management) etc.

The major objectives of CAx technologies are to simplify and to improve human's work (engineer, architect, constructor, physician, surgeon, etc.), by using the hardware and software as an indispensable tool to solve a problem in a certain field (engineering and production, architecture, construction, medicine, economy, business, management, teaching, etc.).

Most commonly used CAx technologies for engineering fields are:

- CAD (Computer-Aided Design) [6-7];
- CADD (Computer-Aided Design and Drafting) [7];
- CAE (Computer-Aided Engineering) [8-14];
- CAM (Computer-Aided Manufacturing) [15-16];
- CAPP (Computer-Aided Process Planning) [17-18];
- CAQ (Computer-Aided Quality) [19];
- CAQ or CAQA (Computer-Aided Quality Assurance) [19];
- CAR (Computer-Aided Robotics) [20];
- CARE (Computer-Aided Reverse Engineering) [4];
- CAS (Computer-Aided Service) [4];
- CAS (Computer-Aided Simulation) [21];

- CASE (Computer-Aided Software Engineering) [22];
- CAT (Computer-Aided Testing) [4];
- CAT (Computer-Aided Tolerancing) [4];
- CIM (Computer-Integrated Manufacturing) [23-24];
- PLM (Product Lifecycle Management) [5] etc.

From mentioned CAX technologies in engineering, the largest number of publications within Scopus have: CAD, CAE, CAM, CAPP, CASE, CIM and PLM technologies, which are also the most famous and most popular CAX technologies. A relatively new field is CAR (computer-aided robotics), which is expected to be widely used in the future.

The most known commercial software tools for CAX technologies are [4, 16]:

- ArchiCAD by Graphisoft, AutoCAD by Autodesk, CATIA by Dassault Systemes, SolidWorks by Dassault Systemes, Pro/ENGINEER (now Creo) by PTC etc., in the field of CAD;
- SolidCAM by Dassault Systemes, FreeCAM; PowerMill by Delcam etc., in the field of CAM;
- CATIA by Dassault Systemes, Creo by PTC, NX by Siemens, PowerShape/PowerMill by Delcam etc., in the field of integrated CAD/CAM;
- Materialise Magics, Netfabb Studio etc., in the field of 3DP (three-dimensional printing), also known as AM (additive manufacturing);
- RapidForm, Geomagic etc., in the field of CARE;
- ANSYS, ABAQUS (Abaqus/Standard and Abaqus/Explicit), Altair, COMSOL Multiphysics, Adams, Moldflow, Moldex3D, 3D-Sigma, LMS Virtual.Lab, LS-DYNA, PAM-CRASH, SAMCEF, RADIOSS etc., in the field of CAE.

### 3. Different Types of Computer-Aided Engineering

Computer-aided engineering (CAE) is a field and technologies that use software systems and tools to solve engineering problems in various scientific fields and industries using various technologies, techniques and methods of engineering analysis [8-14]. CAE is used in all phases of development, design and manufacturing of products, from the initial design to the development phases, to the testing of a product using simulations or rapid prototyping (RP), and finally to planning, manufacturing and quality control of products.

CAE as a term was introduced in the early 1960s. The first publication in the field of CAE technologies in Scopus was published in 1967, and one of the first publications for CAE technologies for engineering applications and industrial production was published in Scopus in the late 1970s [9, 12, 25]. The most cited papers in Scopus where CAE is mentioned only in the "titles of the papers" were reformulated by Zhang et al. in 2003 [14] and by Kwon & Kwon in 2019 [10].

Within the Scopus citation database, there are 16.230 publications or shorter NoP (Number of Publications) for the phrase "computer-aided engineering" (status on the day 09-30-2024) (Figure 2). The maximum NoP for "computer-aided engineering" was in 2005, with 671 publications, and then in 2013, with 633 publications, etc. (Figure 2). In 2022 and 2023, the NoP for CAE technologies decreased compared to the last 30 years, although the development and application of this field in reality did not decrease, but on the contrary increased.

This phenomenon can be explained by the fact that in the "title and abstract of the paper" only the names of software systems and tools, as well as the technologies and methods used in the work are mentioned instead of the word "computer-aided engineering". Likewise, the more intensive development of digital manufacturing and "Industry 4.0" further condition and ensure the increasing application of CAE technologies, which has become an even more important element in the design, testing and production of various products. Also, the experts of the SGR (San Global Research) company in their report dated 28-11-2023 [26] estimate that the global market of CAE technologies was worth 8,4 billion USD in 2022, and predict that it will grow with a value of CAGR (compound annual growth rate) of 11,2% between 2023 and 2032.

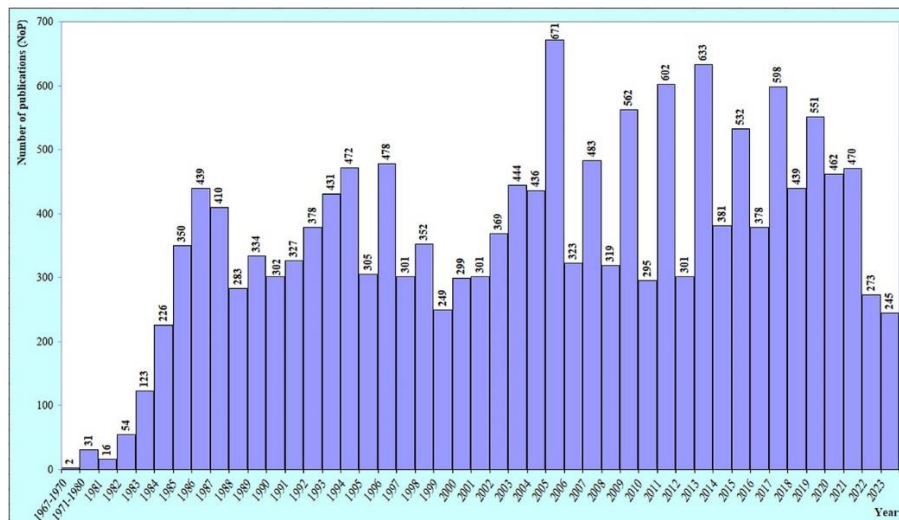


Figure 2: Number of publications (NoP) by year for computer-aided engineering (CAE) in Scopus

For the last 60 years or so, a large number of commercial and free or open source software systems and tools for various engineering analyzes have been developed and used for CAE technologies. Also, standard software systems for CAD technologies that have modules for CAE can also be used for various engineering analyses, such as e.g.: ArchiCAD by Graphisoft, CATIA by Dassault Systems, SolidWorks by Dassault Systems, Pro/ENGINEER (now Creo) by PTC etc.

Also, for the last 60 years or so, a large number of technologies and methods have been developed and used within CAE technologies for engineering analysis within the design, simulation and production of parts, assemblies and complex products in many scientific fields and industrial branches. So the major categories of CAE software systems and tools, for various forms of engineering analysis, include the following technologies and methods: Boundary Element Method (BEM) [27]; Computational Fluid Dynamics (CFD) [28]; Discrete Element Method or also called Distinct Element Method (DEM) [29]; Discontinuity Layout Optimization (DLO) [30]; Finite Difference Method (FDM) [31]; Finite Element Method (FEM) or known as Finite Element Analysis (FEA) [32-34]; Finite Volume Method (FVM) [35]; Infinite Element Method (IEM) [36]; Multibody Dynamics (MBD) [37]; Multidisciplinary Design Optimization (MDO) [38]; Multi-Objective Optimization (MOO) [38]; Thermal Analysis (TA) [34]; Topology Optimization (TO) [39]; Virtual Element Method (VEM) [40] etc., and as a combination of the mentioned methods.

Over time, have been developed several types (variants) of CAE technology, such as:

- CAEBAT (CAE (Computer-Aided Engineering) of Batteries) or (CAE (Computer-Aided Engineering) for Electric Drive Vehicle Batteries) [41];
- CAED (Computer Aided Engineering Design or Computer Aided Engineering and Design) [42-43];
- CCAE (Cloud Computer-Aided Engineering) [33];
- CCAE (Construction Computer-Aided Engineering) [44];
- DCAE (Desktop Computer-Aided Engineering) [45];
- ECAE (Electrical Computer-Aided Engineering) [46];
- ICAE (Integrated Computer-Aided Engineering) [42, 47-48];
- ICAE (Intelligent Computer-Aided Engineering) [49-50];
- ICAED (Integrated Computer-Aided Engineering and Design) [42];
- MCAE (Mechanical Computer-Aided Engineering) [46, 51];
- MCAE (Multidisciplinary Computer-Aided Engineering) [52];
- OCAE (Open-Source Computer-Aided Engineering);
- SaaS-CAE (SaaS (Solver-as-a-Service) with CAE (Computer-Aided Engineering));
- SCAE (Structural Computer-Aided Engineering);
- TCAE (Technological Computer-Aided Engineering);
- TCAE (Total Computer-Aided Engineering) [53];



- WCAE or WBCAE (Web-Based Computer-Aided Engineering) [33];

A large number of systems representing the integration of CAE technologies with other CAX technologies have been developed and implemented. The most common examples of such developed integrations are: CAD/CAE [54-55], CAE/CAD [55], CAD/CAM/CAE [56-57] etc.

CAD and CAE are two of the essential technologies to design and develop new products. CAD/CAE integration can be directed on CAD technologies or it can be directed on CAE technologies when geometric optimization of a product model is required (Figure 3) [55]. Both of these integrations are distinct starting points to integrate with either CAD and CAE systems. However, most of the contemporary research efforts focus on CAD technologies that integrate CAE.

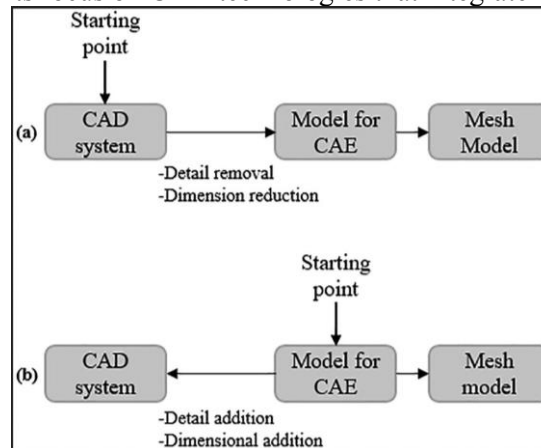


Figure 3: Integrating CAD and CAE: a) Simplified directed on CAD technologies for integrating CAD into CAE; b) Simplified directed on CAE technologies for integrating CAE into CAD [55]

CAD, CAM and CAE are the three basic technologies for designing, developing and manufacturing new products. In CAD/CAM/CAE integration, most current research efforts focus on CAD/CAM technologies that integrate CAE [56-57].

Integrated CAD/CAM/CAE systems offer a wider array of tools and capabilities compared to traditional design methods, empowering organizations to create more intricate and innovative designs. By using CAD/CAM/CAE advanced technologies, companies can create innovative and complex products while securing a competitive advantage. Also, CAD/CAM/CAE systems provide multiple benefits such as costs reduction, enhanced productivity, improved product quality and greater responsiveness to market demands in an industrial environment [57].

## 4. Conclusions

Contemporary design, simulation, testing, manufacturing and quality control of new products is unthinkable without computer-aided technologies or computer-aided everything (CAX) and related technologies. For the field of engineering, these are: CAD, CADD, CAM, CAE, CAPP, CAQ, CAR, CARE, CAS, CASE, CAT, CIM, PLM, etc. A large number of commercial and open source software systems and tools have been developed for all types of CAX technologies to solve engineering problems in various scientific fields and industrial sectors. For design and develop new products two essential technologies are CAD and CAE. For designing, developing and manufacturing new products three basic technologies are CAD, CAM and CAE.

Several different types have been developed CAE technologies (CCAЕ, DCAE, ECAE, ICAE, MCAE, OCAE, SaaS-CAE, SCAE, TCAE, WCAE or WBCAE etc.) and several different forms of integration of CAE technologies with other similar CAX technologies, of which the most commonly used are: CAD/CAE, CAE/CAD, CAD/CAM/CAE, etc.

CAE technologies are also current in the field of publications, so within the Scopus citation database for the phrase "computer-aided engineering" 16.230 NoPs are available (status on the day 30-09-2024).

The global market of CAE technology was valued at USD 8,4 billion in 2022 based on a report by SGR, and is predicted to be worth approximately 12 billion USD in 2026.

**References:**

- [1] J.A. Alic, Computer-assisted everything?: Tools and techniques for design and production. *Technological Forecasting and Social Change*, 44(4) (December 1993), 359–374. ISSN 0040-1625. doi: [10.1016/0040-1625\(93\)90041-5](https://doi.org/10.1016/0040-1625(93)90041-5).
- [2] C.W. Dankwort, R. Weidlich, B. Guenther, J.E. Blaurock, Engineers' CAX education – It's not only CAD. *Computer-Aided Design*, 36(14) (December 2004), 1439–1450. ISSN 0010-4485. doi: [10.1016/j.cad.2004.02.011](https://doi.org/10.1016/j.cad.2004.02.011).
- [3] A.F. De Souza, S. Bodziak, Advanced free form manufacturing by computer aided systems – CAX. In: *Mechanical Engineering. InTech*, Rijeka, Croatia, 2012, pp. 555–586. ISBN 978-953-51-0505-3. doi: [10.5772/36324](https://doi.org/10.5772/36324).
- [4] R. Udriou, Chapter 1: Integration of computer-aided technologies in product lifecycle management (PLM) and human lifecycle management (HUM). In: *Computer-Aided Technologies – Applications in Engineering and Medicine*. IntechOpen, Rijeka, Croatia, 2016, pp. 3–16. ISBN 978-953-51-2788-8. doi: [10.5772/66202](https://doi.org/10.5772/66202).
- [5] Kopei, V.; Onysko, O.; Barz, C.; Dašić, P. & Panchuk, V.: Designing a multi-agent PLM system for threaded connections using the principle of isomorphism of regularities of complex systems. *Machines*, Vol. 11, Issue 2 (February 2023), Article no. 263: pp. 1–22. eISSN 2075-1702. doi: [10.3390/machines11020263](https://doi.org/10.3390/machines11020263).
- [6] P. Harrow, Factors in selecting a computer-aided design system. *Computer-Aided Engineering Journal*, 1(1) (November 1983), 16–24. ISSN 0263-9327. doi: [10.1049/cae.1983.0006](https://doi.org/10.1049/cae.1983.0006).
- [7] A. Molga, M. Wójtowicz, Computer-aided design systems. *Edukacja-Technika-Informatyka*, 3(2) (2012), 291–297. ISSN 2080-9069.
- [8] M. Benayoune, P.E. Preece, Review of information management in computer-aided engineering. *Computers and Chemical Engineering*, 11(1) (1987), 1–6. ISSN 0098-1354. doi: [10.1016/0098-1354\(87\)80001-7](https://doi.org/10.1016/0098-1354(87)80001-7).
- [9] R.S. Davies, A. Clarke, Computer aided engineering for manufacturing industries. *Chartered Mechanical Engineer*, Vol. 26, Issue 11 (December 1979), pp. 43–48. ISSN 0009-191X.
- [10] H.-J. Kwon, H.-K. Kwon, Computer aided engineering (CAE) simulation for the design optimization of gate system on high pressure die casting (HPDC) process. *Robotics and Computer-Integrated Manufacturing*, 55: Part B (February 2019), 147–153. ISSN 0736-5845. doi: [10.1016/j.rcim.2018.01.003](https://doi.org/10.1016/j.rcim.2018.01.003).
- [11] D.-C. Lee, C.-S. Han, CAE (computer aided engineering) driven durability model verification for the automotive structure development. *Finite Elements in Analysis and Design*, 45(5) (April 2009), 324–332. ISSN 0168-874X. doi: [10.1016/j.finel.2008.10.004](https://doi.org/10.1016/j.finel.2008.10.004).
- [12] J.Q. Parmater, Computer-aided engineering: The step beyond CAD/CAM. *Machine Design*, 52(24) (October 1980), pp. 55–59. ISSN 0024-9114.
- [13] B. Raphael, I.F.C. Smith, *Engineering Informatics: Fundamentals of computer-aided engineering*. 2nd Edition. John Wiley & Sons Ltd., Chichester, West Sussex, United Kingdom, 2013. – 400 pp. ISBN 978-1-119-95341-8.
- [14] G.P. Zhang, Y.M. Huang, W.H. Shi, W.P. Fu, Predicting dynamic behaviours of a whole machine tool structure based on computer-aided engineering. *International Journal of Machine Tools and Manufacture*, 43(7) (May 2003), pp. 699–706. ISSN 0890-6955. doi: [10.1016/S0890-6955\(03\)00026-9](https://doi.org/10.1016/S0890-6955(03)00026-9).
- [15] T.C. Chang, R.A. Wysk, H.P. Wang, *Computer aided manufacturing*. Prentice Hall, Upper Saddle River, New Jersey, 1998.
- [16] P. Dašić, Scientific and technological trends: Selected scientific-professional papers (in Serbian). SaTCIP Publisher Ltd., Vrnjačka Banja, 2020. – 305 str. ISBN 978-86-6075-072-5.
- [17] L. Alting, H.-C. Zhang, Computer aided process planning: The state-of-the-art survey. *International Journal of Production Research*, 27(4) (April 1989), pp. 553–585. ISSN 0020-7543. doi: [10.1080/00207548908942569](https://doi.org/10.1080/00207548908942569).
- [18] Y. Yusof, K. Latif, Survey on computer-aided process planning. *The International Journal of Advanced Manufacturing Technology*, 75(1-4) (October 2014), pp. 77–89. ISSN 0268-3768. doi: [10.1007/s00170-014-6073-3](https://doi.org/10.1007/s00170-014-6073-3).

- [19] M. Innala, S. Torvinen, Integrating computer aided quality assurance in flexible engineering and production processes. *International Journal of Production Economics*, 41 (1995), 179–184. ISSN 0925-5273.
- [20] P. Neto, N. Mendes, R. Araújo, J.N. Pires, High-level robot programming based on CAD: Dealing with unpredictable environments. *Industrial Robot: An International Journal*, 39(3) (2012), 294–303. ISSN 0143-991X. doi: [10.1108/01439911211217125](https://doi.org/10.1108/01439911211217125).
- [21] M.S. Sidhu, K. Ali, Computer aided simulation of a 4BL engineering problem using Matlab. *Australian Journal of Basic and Applied Sciences*, 7(10) (2013), 381–387. ISSN 1991-8178.
- [22] R. Leggate, Computer-aided software engineering – Turning the corner. *Computer-Aided Engineering Journal*, 1(1) (November 1983), 9–15. ISSN 0263-9327. doi: [10.1049/cae.1983.0005](https://doi.org/10.1049/cae.1983.0005).
- [23] D.K. Allen, J. Bryan, Architecture for computer-integrated manufacturing. *CIRP Annals – Manufacturing Technology*, 35(1) (June 1986), pp. 351–354. ISSN 0007-8506. doi: [10.1016/S0007-8506\(07\)61904-0](https://doi.org/10.1016/S0007-8506(07)61904-0).
- [24] J.-B. Waldner, CIM: Principles of computer-integrated manufacturing. John Wiley & Sons Ltd., Chichester, West Sussex, United Kingdom, 1992. – 206 pp. ISBN 978-0-471-93450-9.
- [25] Jones, M.V.: The documentation and checking of computer aided engineering computations. *Computers and Structures*, 10(1-2) (April 1979), 81–84. ISSN 0045-7949. doi: [10.1016/0045-7949\(79\)90075-0](https://doi.org/10.1016/0045-7949(79)90075-0).
- [26] Computer Aided Engineering Market – Global industry, analysis, size, share, growth, trends, and forecasts 2024-2032. Report ID 3142. San Global Research (SGR), Pune, India, 2023. – 150 pp.
- [27] L.C. Wrobel, The boundary element method, Vol. 1: Applications in thermo-fluids and acoustics. John Wiley & Sons Ltd., Chichester, West Sussex, United Kingdom, 2002. – 472 pp. ISBN 978-0-471-72039-3.
- [28] S. Mahmood, D. Huang, Computational fluid dynamics based bulbous bow optimization using a genetic algorithm. *Journal of Marine Science and Application*, 11(3) (September 2012), 286–294. ISSN 1671-9433. doi: [10.1007/s11804-012-1134-1](https://doi.org/10.1007/s11804-012-1134-1).
- [29] Z.-B. Peng, E. Doroodchi, B. Moghtaderi, Heat transfer modelling in discrete element method (DEM)-based simulations of thermal processes: Theory and model development. *Progress in Energy and Combustion Science*, 79 (July 2020), Article no. 100847. ISSN 0360-1285. doi: [10.1016/j.pecs.2020.100847](https://doi.org/10.1016/j.pecs.2020.100847).
- [30] C. Smith, M. Gilbert, Application of discontinuity layout optimization to plane plasticity problems. *Proceedings of the Royal Society, A: Mathematical, Physical and Engineering Sciences*, 463(2086), (October 2007), 2461–2484. ISSN 1364-5021. doi: [10.1098/rspa.2006.1788](https://doi.org/10.1098/rspa.2006.1788).
- [31] A.R. Mitchell, D.F. Griffiths, The finite difference method in partial differential equations. John Wiley & Sons Ltd., Chichester, West Sussex, United Kingdom, 1980. – 272 pp.
- [32] E. Desnica, M. Đurđev, B. Vašćić, R. Turmanidze, P. Dašić, Determination of a safety factor of a car wheel rim using finite element analysis in SolidWorks. *Applied Engineering Letters*, 7(4) (December 2022), 163–171. ISSN 2466-4677. doi: [10.18485/aeletters.2022.7.4.4](https://doi.org/10.18485/aeletters.2022.7.4.4).
- [33] Y. Ihara, G. Hashimoto, H. Okuda, Web-based integrated cloud CAE platform for large-scale finite element analysis. *Mechanical Engineering Letters*, 3 (2017), Article no. 17-00520: pp. 1–8. ISSN 2189-5236. doi: [10.1299/mel.17-00520](https://doi.org/10.1299/mel.17-00520).
- [34] O.C. Zienkiewicz, R.L. Taylor, J.Z. Zhu, The finite element method: Its basis and fundamentals. 7th Edition. Butterworth-Heinemann & Elsevier Ltd., Oxford, United Kingdom, 2013. – 756 pp. ISBN 978-1-85617-633-0.
- [35] Z.-C. Li, S. Wang, The finite volume method and application in combinations. *Journal of Computational and Applied Mathematics*, 106(1) (June 1999), 21–53. ISSN 0377-0427. doi: [10.1016/S0377-0427\(99\)00051-5](https://doi.org/10.1016/S0377-0427(99)00051-5).
- [36] L.A. Ying, An introduction to the infinite element method (in Chinese). *Mathematics in Practice and Theory*, 22(2) (1992), 69–78. ISSN 1000-0984.
- [37] C.M. Pappalardo, D. Guida, On the use of two-dimensional Euler parameters for the dynamic simulation of planar rigid multibody systems. *Archive of Applied Mechanics*, 87(10) (October 2017), 1647–1665. ISSN 0939-1533. doi: [10.1007/s00419-017-1279-0](https://doi.org/10.1007/s00419-017-1279-0).
- [38] R. Ghadge, R. Ghorpade, S. Joshi, Multi-disciplinary design optimization of composite structures: A review. *Composite Structures*, 280 (January 2022), Article no. 114875. ISSN 0263-8223. doi: [10.1016/j.compstruct.2021.114875](https://doi.org/10.1016/j.compstruct.2021.114875).

- [39] X. Yan, X. Huang, Y. Zha, Y.-M. Xie, Concurrent topology optimization of structures and their composite microstructures. *Computers & Structures*, 133 (March 2014), 103–110. ISSN 0045-7949. doi: [10.1016/j.compstruc.2013.12.001](https://doi.org/10.1016/j.compstruc.2013.12.001).
- [40] L. Beirão Da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L.D. Marini, A. Russo, Basic principles of virtual element methods. *Mathematical Models and Methods in Applied Sciences*, 23(1) (January 2013), 199–214. ISSN 0218-2025. doi: [10.1142/S0218202512500492](https://doi.org/10.1142/S0218202512500492).
- [41] S. Santhanagopalan, (editor), *Computer aided engineering of batteries*. Springer Nature Switzerland AG, Cham, Switzerland, 2023. – 280 pp. ISSN 0076-9924 and ISBN 978-3-031-17606-7. doi: [10.1007/978-3-031-17607-4](https://doi.org/10.1007/978-3-031-17607-4).
- [42] C.-M. Huang, A. Kucinic, J.A. Johnson, H.-J. Su, C.E. Castro, Integrated computer-aided engineering and design for DNA assemblies. *Natural Materials*, Vol. 20(9) (September 2021), 1264–1271. ISSN 1476-1122. doi: [10.1038/s41563-021-00978-5](https://doi.org/10.1038/s41563-021-00978-5).
- [43] A. Saxena, B. Sahay, *Computer aided engineering design*. Springer Science+Business Media Inc., New York, USA, 2005. – 394 pp. ISBN 978-1-4020-2555-6. doi: [10.1007/1-4020-3871-2](https://doi.org/10.1007/1-4020-3871-2).
- [44] G. Belonogoff, A. Aguilar, F. Aguilar, Developing integrated technology systems. *Nuclear Plant Journal*, 12(4) (July 1994), 16–18. ISSN 0892-2055.
- [45] L. Reifschneider, Teaching design for manufacturability with desktop computer-aided analysis. *Journal of Industrial Technology*, 16(3) (May-July 2000), 2–5.
- [46] Garfinkel, G.A.: Integrating mechanical computer-aided engineering tools into the printed wiring board design cycle. In: *Proceedings of the 6th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM'98)*; Seattle, WA, USA; 27-30 May 1998. Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ (New Jersey), USA, 1998, pp. 316–321. ISSN 1089-9870 and ISBN 0-7803-4475-8. doi: [10.1109/ITHERM.1998.689580](https://doi.org/10.1109/ITHERM.1998.689580).
- [47] D.W. Douglas, Integrated computer-aided engineering (ICAE), Part I – Systems design. In: *Proceedings of the 6th Annual Conference & Exposition on Computer Graphics'85*. National Computer Graphics Association (NCGA), Fairfax, Virginia, USA, 1985, pp. 47–61.
- [48] R.A. Peters, Integrated computer-aided engineering. *Chemical Engineering*, 92(10) (1985), pp. 95–96, 98. ISSN 0009-2460.
- [49] K.D. Forbus, Intelligent computer-aided engineering. *AI Magazine*, 9(3) (Fall 1988), 23–36. ISSN 0738-4602. doi: [10.1609/aimag.v9i3.940](https://doi.org/10.1609/aimag.v9i3.940).
- [50] W.-B. Hou, H.-Z. Zhang, R.-F. Chi, P. Hu, Development of an intelligent CAE system for auto-body concept design. *International Journal of Automotive Technology*, 10(2) (April 2009), 175–180. ISSN 1229-9138. doi: [10.1007/s12239-009-0021-5](https://doi.org/10.1007/s12239-009-0021-5).
- [51] R.C. Miller, Setting the standards for desktop MCAE systems. *Design News*, 43(2) (January 1987), 90–92, 94. ISSN 0011-9407.
- [52] T.E. Fenske, Computer-aided structural engineering software: The status of concurrent engineering. In: *Proceedings of the 11th Conference on Analysis and Computation*; Atlanta, Georgia, USA; 24-28 April 1994. American Society of Civil Engineers (ASCE), New York, 1994, pp. 230–237. ISBN 978-0-87262-974-5.
- [53] S. Romero-Hernández, O. Romero, A framework of computer aided engineering and LCA applied for life cycle management life. In: *Proceedings of the International Life Cycle Assessment and Management Conference (InLCA/LCM-2003)*; Seattle, Washington, USA; 22-25 September 2003.
- [54] K.-H. Chang, *Design theory and methods using CAD/CAE*. Academic Press Inc. & Elsevier Inc., San Diego, California, USA, 2015. – 516 pp. ISBN 978-0-12-398512-5.
- [55] M.T.H. Khan, S. Rezwana, A review of CAD to CAE integration with a hierarchical data format (HDF)-based solution. *Journal of King Saud University – Engineering Sciences*, 33(4) (May 2021), 248–258. ISSN 1018-3639. doi: [10.1016/j.jksues.2020.04.009](https://doi.org/10.1016/j.jksues.2020.04.009).
- [56] K. Lee, *Principles of CAD/CAM/CAE systems*. Addison Wesley Longman Inc., Reading, Massachusetts, USA, 1999. – 582 pp. ISBN 978-0-201-38036-1.
- [57] A.M. Țițu, A.B. Pop, Implementation of CAD/CAM/CAE systems for improved design and manufacturing processes in industrial organizations. *Proceedings of the 18th International Conference on Business Excellence (PICBE)*, 18(1) (June 2024), 3069–3078. ISSN 2558-9652. doi: [10.2478/picbe-2024-0253](https://doi.org/10.2478/picbe-2024-0253).

# Communication Challenges in Agile Meetings

Maja Gaborov<sup>1</sup>, Miodrag Kovačević<sup>2</sup>, Nemanja Tasić<sup>1</sup>, Dragan Kreculj<sup>3</sup>, Nada Ratković Kovačević<sup>3</sup> and Dragana Kovač<sup>1</sup>

<sup>1</sup>University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

<sup>2</sup>Technical College of Applied Sciences, Zrenjanin, Serbia

<sup>3</sup>The Academy of Applied Studies Polytechnic, New Belgrade Section, Belgrade, Serbia

maja.gaborov@tfzr.rs; miodrag.kovacevic.zr@gmail.com; ntasic89@gmail.com;  
dkreculj@politehnika.edu.rs; nratkovickovacevic@politehnika.edu.rs, dragana.milosavljev@tfzr.rs

## Abstract:

This paper explores theoretical aspects of communication challenges in agile meetings, which are critical to agile project management. By analyzing existing literature on communication dynamics, meeting effectiveness, and the unique challenges of virtual environments, this study aims to identify key communication issues and propose theoretical solutions to enhance meeting productivity and team cohesion. The descriptions and consequences of the most important communication challenges in agile meetings that impede collaboration and cooperation are listed in the literature. The results indicate the importance of solving communication problems in agile meetings. The paper also suggests solutions for improving communication in agile meetings.

## Keywords:

Communication, challenges, agile, meetings

## 1. Introduction

Agile methodologies prioritize frequent and structured meetings to facilitate continuous improvement and collaboration among team members. However, communication challenges can significantly impact the effectiveness of these meetings.

Communication in agile meetings plays a key role in ensuring the success of teamwork and the achievement of project goals [1]. It is essential because it allows team members to regularly share information regarding progress, obstacles and next steps [2].

Different types of agile meetings are [2]:

- Daily Stand-up: A short, daily team meeting for members to share what they did the previous day, what they plan to do today, and if they encountered any obstacles.
- Sprint Planning: A meeting at the beginning of each sprint where the team agrees on goals for that sprint, prioritizes tasks, and assesses what can be accomplished.
- Sprint Review: A meeting at the end of a sprint where the team and stake-holders review and evaluate what has been accomplished, usually including a demonstration of completed work units.
- Sprint Retrospective: A meeting at the end of a sprint where the team analyzes what worked well, what can be improved, and plans concrete steps to implement improvements.
- Backlog Refinement: A meeting where backlog tasks are elaborated and coordinated, requirements are clarified and prioritization is planned for future sprints.
- Ad hoc meetings: In addition to standard agile meetings, teams can meet ad hoc to solve urgent problems or unplanned situations that require quick intervention.

Each of these meetings has its own purpose in the agile methodology and contributes to better project management and team productivity. This paper investigates common communication problems in agile meetings and their implications for team dynamics and productivity. The aim of this paper is to show all communication problems in agile meetings, to find out what the consequences are and to present potential solutions for them.

## 2. Related work

In the previous work [3], positive and negative items that affect the success of the project were shown, where the focus was placed on reducing job dissatisfaction and motivating team members to be even more satisfied.

The authors argue that frequent and open communication helps in the early identification of potential problems, thus enabling timely interventions and fostering a culture of continuous improvement [4].

The need for open, frequent, and transparent communication is emphasized to respond to challenges and ensure the alignment of large teams [5].

The book [6] provides an overview of agile methodologies and emphasizes the importance of communication in agile teams. It covers best practices for effective communication in various agile frameworks, including Scrum and Kanban.

## 3. Communication dynamics in agile meetings

Agile methodologies rely heavily on effective communication to ensure that all team members are aligned and can collaborate efficiently [2]. Regular meetings, such as daily stand-ups, sprint planning, and retrospectives, are designed to enhance transparency and foster open dialogue [7].

Frequent meetings can lead to information overload, where team members struggle to process and retain large amounts of information [8]. Not all team members may have equal opportunities to contribute, leading to the dominance by a few voices and the marginalization of others [9-11].

## 4. Communication barriers and meeting effectiveness

Several barriers can impede effective communication in agile meetings, including cultural differences, hierarchical dynamics, and unclear communication channels [12].

Cultural Barriers mean that teams may face challenges related to different communication styles and expectations, potentially leading to misunderstandings and conflicts [13]. Power imbalances within the team can discourage open communication and honest feedback, negatively impacting team cohesion and decision-making [14].

Effective meeting structures are essential for facilitating clear and productive communication. Meetings with well-defined agendas, time management, and facilitation practices can help mitigate communication barriers [11-12].

## 5. Virtual meetings and communication challenges

The rise of virtual meetings, especially in the IT (Information Technologies) sector, introduces unique communication challenges. The lack of nonverbal cues, increased cognitive load, and technological issues can hinder effective communication [17-18].

The absence of physical presence in virtual meetings reduces the ability to interpret body language and other nonverbal signals, which are crucial for effective communication [19].

Virtual meetings require [20]: additional cognitive effort to interpret digital cues and maintain attention, leading to "Zoom fatigue" and decreased communication effectiveness.

## 6. Discussion

Agile methodologies rely on effective communication to ensure that all team members are aligned and can collaborate effectively [2]. Regular meetings such as daily stand-ups, sprint planning, and retrospectives are designed to improve transparency and encourage open dialogue [7]. In discussing communication dynamics in agile meetings, key points to consider include the importance of communication in agile practices, identification of key challenges such as information overload and

unequal participation, barriers to effective communication such as cultural differences and hierarchical dynamics, the impact of virtual environments on communication processes, theoretical and practical implications of your findings, as well as future research directions. Together, these aspects provide a deeper understanding of how communication affects the effectiveness of agile teams and offer guidelines for improving practices in agile project management.

Information overload is a common problem, where team members may have difficulty processing large amounts of information during frequent meetings [8].

Also, unequal participation can result in the dominance of certain voices and the marginalization of other team members [9-11].

In agile meetings, barriers to effective communication include cultural differences, hierarchical dynamics, and unclear communication channels [12]. The structure of meetings with clearly defined agendas, time management and effective facilitation is also important in order to overcome these barriers [15-16].

In the context of virtual environments, challenges such as a lack of nonverbal cues, increased cognitive effort, and technological problems can significantly impede effective communication [17-18]. Using visual aids, collaborative tools, and regular breaks can be key to improving communication in virtual meetings [18], [20].

In Table 1 are listed descriptions and consequences of the most important communication challenges in Agile Meetings which impede collaboration and cooperation, with referencing sources.

**Table 1:**  
Communication Challenges in Agile Meetings: Descriptions and Consequences

Communication Problem	Description	Consequences	References
Information Overload	Frequent meetings lead to excessive information intake.	Reduced ability to focus, increased stress, information overload.	[8]
Unequal Participation	Discussions dominated by some members, while others contribute less.	Marginalization of less active members, lack of diversity in ideas.	[9-11]
Cultural Differences	Cultural diversity affects communication styles and message interpretation.	Misunderstandings, conflicts, reduced team efficiency.	[12]
Hierarchical Dynamics	Hierarchical structures inhibit open idea exchange and communication.	Lack of honest feedback, limitation of innovative solutions.	[15-16]
Lack of Nonverbal Cues	Virtual meetings lack nonverbal signals interpretation.	Reduced ability to understand tones and emotional nuances.	[17-18]
Increased Cognitive Load	Virtual environments require additional effort to process information.	"Zoom fatigue," decreased concentration, lower efficiency in long meetings.	[18], [20-21]
Technological Issues	Technology problems (connectivity, applications) disrupt communication.	Communication breakdowns, loss of important information, participant frustration.	[17-18]

## 7. Research implications

Understanding communication challenges in agile meetings illuminates key aspects of teamwork effectiveness and provides a basis for practical and theoretical implications that can improve meeting management strategies.

### 7.1. Practical Implications

- **Enhancing Meeting Effectiveness:** Understanding communication challenges in agile meetings can lead to the development of better strategies for meeting management. For instance, establishing clear agendas and ensuring equitable participation can make meetings more efficient and reduce cognitive fatigue among team members [15-16], [9-11]. This insight can guide managers in structuring meetings to maximize productivity and engagement.
- **Improving Virtual Communication:** Given the increasing prevalence of virtual meetings in the IT sector, identifying the unique challenges of these settings is crucial. The insights from this study suggest that incorporating visual aids and scheduling regular breaks can mitigate the effects of "Zoom fatigue" and enhance communication effectiveness [18], [21]. This can inform best practices for remote team management and improve overall team performance.
- **Addressing Cultural and Hierarchical Barriers:** The findings highlight the importance of addressing cultural and hierarchical barriers to communication. By promoting psychological safety and fostering an inclusive environment, teams can enhance open dialogue and collaboration [14], [12]. This can be particularly beneficial in diverse and multinational teams, where cultural differences may pose significant communication challenges.

### 7.2. Theoretical Implications

This research aims to find important communication challenges in agile meetings as well as to propose potential solutions. This study contributes to the body of knowledge on agile methodologies by providing a detailed examination of communication issues in agile meetings. It extends existing theories on meeting effectiveness and cognitive load by applying them to the agile context [2], [22]. Future research can build on these findings to explore other aspects of agile practices and their impact on team dynamics.

## 8. Solution review for improving communication in agile meetings

Several solutions have been proposed to improve communication in agile meetings: clear and detailed agendas to ensure focused discussion and efficient time management; facilitation techniques that encourage equal participation of all team members; cultural awareness training to reduce misunderstandings and conflicts; promoting psychological safety for open dialogue and building trust; using visual tools in virtual meetings for better understanding and engagement; and regular breaks during extended virtual meetings to reduce cognitive load and fatigue.

Suggested solutions to reduce communication problems in agile meetings include:

- **Introducing clearly defined agendas:** Establishing detailed agendas for each meeting to ensure focused discussion and efficient time management [9].
- **Encouraging equal participation:** Use of facilitation techniques that encourage equal participation of all team members, thereby ensuring that all perspectives are heard and valued [15-16].
- **Cultural Awareness Training:** Educating teams about cultural differences in communication to reduce misunderstandings and conflicts [13].
- **Creating an inclusive environment [14]:** Promoting psychological safety within teams to encourage open dialogue and build trust.
- **Using visual aids in virtual meetings:** Integrating visual tools and collaborative platforms to facilitate understanding and engagement in virtual meetings [18].



- Frequent breaks during virtual meetings: Regularly include breaks during extended virtual meetings to reduce cognitive strain and fatigue [22].

A Company can derive/ build the Corporate Communication Strategy and offer it as a service to a customer company. An example of good practice is [23]. Services may include [23]: stakeholder communication planning; crisis communication management; thought leadership and content marketing; internal communication and change management.

Proposed solutions and description of solutions are presented in Table 2.

**Table 2:**  
Solutions for Improving Communication in Agile Meetings

<b>Solution</b>	<b>Description</b>	<b>Source</b>
Clearly Defined Agendas	Establishing detailed agendas for each meeting to ensure focused discussion and efficient time management.	[9]
Encouraging Equal Participation	Use of facilitation techniques that encourage equal participation of all team members, thereby ensuring that all perspectives are heard and valued.	[15-16]
Cultural Awareness Training	Educating teams about cultural differences in communication to reduce misunderstandings and conflicts.	[13]
Creating an Inclusive Environment	Promoting psychological safety within teams to encourage open dialogue and build trust.	[14]
Using Visual Aids in Virtual Meetings	Integrating visual tools and collaborative platforms to facilitate understanding and engagement in virtual meetings.	[18]
Frequent Breaks During Virtual Meetings	Regularly including breaks during extended virtual meetings to reduce cognitive strain and fatigue.	[22]

## 9. Conclusions

Effective communication is crucial for the success of agile meetings. By understanding and addressing the common communication challenges identified in this paper, agile teams can improve meeting effectiveness, enhance team cohesion, and boost overall productivity.

In this paper communication barriers and the efficiency of the meeting are presented. Then virtual meetings related to communication were analyzed in detail. Also, descriptions and consequences of the most important communication challenges in agile meetings have been listed. This paper identifies key communication problems in agile meetings and their consequences. Information overload leads to reduced ability to focus and increased stress, while unequal participation results in less activity among team members and less diversity of ideas. Cultural differences cause misunderstandings and conflicts, and hierarchical dynamics limit the open exchange of ideas and innovation. The lack of non-verbal signals in virtual meetings makes it difficult to understand tones and emotions, while the increased cognitive load causes fatigue and reduced efficiency. Ultimately, technological problems lead to communication breakdowns and frustration among participants. Moreover, specific solutions to improve communication in agile meeting scenarios have been discussed. The paper proposes many solutions for improving communications in agile meetings. Introducing detailed agendas for each meeting allows for focused discussion and efficient time management. The use of facilitation techniques encourages the equal participation of all members, ensuring that all perspectives are heard and valued. Education about cultural differences helps reduce misunderstandings and conflicts in multicultural teams while promoting psychological safety encouraging open dialogue and building

trust. The integration of visual tools and collaborative platforms improves understanding and engagement in virtual meetings, and regular breaks during long virtual meetings reduce workload and fatigue.

In the future, an empirical survey of IT experts will be conducted in order to discover more problems that may arise during communication in agile meetings. A software tool will be created that can be used to facilitate communication in meetings.

#### **Acknowledgment:**

The Authors extend their gratitude towards their affiliating higher education institutions for support in this work.

#### **References:**

- [1] J. A. Highsmith III, *Adaptive Software Development: A Collaborative Approach to Managing Complex Systems*, 1<sup>st</sup> ed., Dorset House Publishing Co., Inc. New York, NY USA, 2001.
- [2] K. Schwaber, and J. Sutherland, *The Scrum Guide™ - The Definitive Guide to Scrum: The Rules of the Game*, 2017. URL: <https://scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf>.
- [3] M. Gaborov, D. Kovač, N. Tasić, D. Kreculj, and N. Ratković Kovačević, Job Satisfaction in Scrum Teams, in: *Proceedings of the 5<sup>th</sup> Virtual International Conference on Path to a Knowledge Society-Managing Risks and Innovation, PaKSoM 2023*, Mathematical Institute SANU/ SASA, Belgrade, Serbia, 2023, pp. 335-339.
- [4] L. M. Maruping, V. Venkatesh, and R. Agarwal, A control theory perspective on agile methodology use and changing user requirements, *Information Systems Research* 20(3) (2009) 377-399. doi: <https://doi.org/10.1287/isre.1090.0238>
- [5] D. K. Rigby, S. Berez, and A. Noble, "Agile at scale." *Harvard Business Review* 5(2018) (May-June 2018): 88-96. URL: <https://hbr.org/2018/05/agile-at-scale>
- [6] P. Measey, and R. Rattac, *Agile foundations: Principles, practices, and frameworks*, reprinted ed., BCS Learning & Development Limited - a wholly owned subsidiary of BCS, The Chartered Institute for IT, Swindon, United Kingdom, 2015.
- [7] A. Cockburn, and J. Highsmith, "Agile software development: The people factor." *Computer* 34(11) (2001): 131-133. doi: <https://doi.org/10.1109/2.963450>
- [8] M. J. Eppler, and J. Mengis, "The Concept of Information Overload: A Review of Literature From Organization Science, Accounting, Marketing, MIS, and Related Disciplines." *The Information Society* 20(5) (2004): 325-344. doi: <https://doi.org/10.1080/01972240490507974>
- [9] N. K. Lehmann-Willenbrock, J. A. Allen, and S. Kauffeld, "A sequential analysis of procedural meeting communication: How teams facilitate their meetings." *Journal of Applied Communication Research* 41(4) (2013): 365-388. URL: <https://research.vu.nl/ws/portalfiles/portal/833560/L-W%2C+Allen%2C+Kauffeld.2013.pdf> doi: <https://doi.org/10.1080/00909882.2013.844847>
- [10] N. K. Lehmann-Willenbrock, and J. A. Allen, "How fun are your meetings? Investigating the relationship between humor patterns in team interactions and team performance." *Journal of Applied Psychology* 99(6) (2014): 1278-1287. doi: <https://doi.org/10.1037/a0038083> URL: <https://research.vu.nl/ws/portalfiles/portal/1118344/Lehmann-Willenbrock%2BAllen+2014+How+fun+are+your+meetings.pdf>
- [11] J. A. Allen, N. K. Lehmann-Willenbrock, and S. G. Rogelberg (Eds.), *The Cambridge Handbook of Meeting Science*. Cambridge Handbooks in Psychology, Cambridge University Press, Cambridge, 2015. doi: <https://doi.org/10.1017/CBO9781107589735>
- [12] J. M. Brett, K. J. Behfar, and M. C. Kern, Managing multicultural teams. *Harvard Business Review* 84(11) (Nov. 2006): 84-91. URL: <https://hbr.org/2006/11/managing-multicultural-teams>
- [13] S. Ting-Toomey, and T. Dorjee, *Communicating Across Cultures*, 2nd. ed., The Guilford Press – A Division of Guilford Publications, Inc., New York & London, NY USA & UK, 2019.
- [14] A. C. Edmondson, "Psychological Safety and Learning Behavior in Work Teams." *Administrative Science Quarterly* 44(2) (1999): 350-383. doi: <https://doi.org/10.2307/2666999>

- [15] S. G. Rogelberg, J. A. Allen, L. Shanock, C. Scott, and M. Shuffler. "Employee satisfaction with meetings: A contemporary facet of job satisfaction." *Human Resource Management* 49(2) (2010): 149-172. doi: <https://doi.org/10.1002/hrm.20339>
- [16] S. G. Rogelberg, D. J. Leach, P. B. Warr, and J. L. Burnfield. "Not another meeting! Are meeting time demands related to employee well-being?" *Journal of Applied Psychology* 91(1) (2006): 83-96. doi: <https://doi.org/10.1037/0021-9010.91.1.83>
- [17] R. L. Daft, and R. H. Lengel. "Organizational Information Requirements, Media Richness and Structural Design." *Management Science* 32(5) (1986): 554-571. doi: <https://doi.org/10.1287/mnsc.32.5.554>
- [18] A. R. Dennis, and J. S. Valacich, Rethinking Media Richness: Towards a Theory of Media Synchronicity, in: *Proceedings of the 32nd Hawaii International Conference on System Sciences*, 1999, HICSS-32, Abstracts and CD-ROM of Full Papers, Maui, HI, USA, (1999): 1-10. URL: <https://collablab.northwestern.edu/CollablabDistro/nucmc/DennisAndValacich-RethinkingMediaRichness-HICSS-1999.pdf> doi: <https://doi.org/10.1109/HICSS.1999.772701>.
- [19] A. Mehrabian, *Silent Messages*, 1<sup>st</sup> ed., Wadsworth Publishing Company, Inc., Belmont, CA, USA, 1971.
- [20] L. Fosslien, and M. W. Duffy, How to combat Zoom fatigue, *Harvard Business Review*, Apr. 2020. URL: <https://hbr.org/2020/04/how-to-combat-zoom-fatigue>
- [21] G. Mark, D. Gudith, and U. Klocke, The cost of interrupted work: More speed and stress, in: *Proceedings of the SIGCHI International Conference on Human Factors in Computing Systems*, SIG CHI '08, ACM, New York, NY, USA, 2008, pp. 107-110. URL: <https://api.semanticscholar.org/CorpusID:207167051> doi: <https://doi.org/10.1145/1357054.1357072>
- [22] J. Sweller, Cognitive load during problem solving: Effects on learning, *Cognitive Science* 12(2) (1988): 257-285. doi: [https://doi.org/10.1207/s15516709cog1202\\_4](https://doi.org/10.1207/s15516709cog1202_4)
- [23] Merkle Innovation, Consulting - Corporate Communication Strategy, URL: <https://merkleinnovation.com/corporate-communications-strategy/>

# Analysis of GPT-4o Application in Solving Simple Static Beams Based on Uploaded Images: Potentials and Limitations

Ljubica Lazić Vulićević<sup>1</sup>, Željko Eremić<sup>1</sup> and Valentina Mladenović<sup>1</sup>

<sup>1</sup> Technical College of Applied Sciences in Zrenjanin, Djordja Stratimirovica 23, Zrenjanin, Serbia

lazic.ljubica@yahoo.com; zeljko.eremic@gmail.com; mladenovic.val@gmail.com

## Abstract:

This paper explores the potential of GPT-4o for use in solving statics problems through the analysis of static structures, specifically simple beam, based on uploaded images. By leveraging image recognition technologies, GPT-4o can identify and interpret key elements of static systems, such as forces, moments, supports, and loads, and apply fundamental principles, like force and moment equilibrium. Its application in education is promising, allowing students to efficiently comprehend basic structural analyses without manual calculation. Similarly, in engineering practice, it could serve as a tool for preliminary analyses of simpler systems. This study aims to evaluate GPT-4o's practicality, reliability, and potential in the educational process. Results show that while GPT-4o successfully identifies loads and analyzes structures using basic statics principles, limitations exist, particularly in its ability to interpret results accurately.

## Keywords:

artificial intelligence, GPT-4o, statics beams, image recognition, engineering education, educational tool.

## 1. Introduction

GPT-4o has the potential to be used in various engineering fields, including, among others, solving statics problems, including the analysis of static structures, based on an uploaded image or diagram. GPT-4 can utilize technologies for recognizing and understanding visual data to interpret solutions for static structures, such as beam supports, cantilever beams, or frames. Uploaded images can contain sketches with forces, moments, supports, and loads, and the software can identify these elements and their role in the system. Based on the uploaded image, AI can theoretically analyze a static system using the basic principles of statics, such as the equilibrium of forces and moments. In the case of beam supports, this means identifying the forces acting on the beam and calculating support reactions, bending moments, shear forces, and potential displacements. Although GPT-4o can interpret basic tasks in statics, there are currently limitations in these analyses, which will be discussed further in this paper.

One of the greatest potentials of this technology could be its use in the educational process, where students can quickly gain insights into basic analyses of static structures without the need for manual calculations. Additionally, in engineering practice, it could be used for rapid, preliminary analysis of simpler systems before the task is handed over to specialized software.

The idea presented in this paper is to test GPT-4o in solving simple statics problems related to beam supports, and later, more complex problems as well. The aim was to determine how practical this new capability offered by the software is, how reliable the results are, and whether it could be used in the educational process to assist students in solving problems or verifying results obtained through calculations. The results show that the software effectively demonstrates the problem-solving process, correctly identifies the loads acting on the structure, and analyzes the static system using the basic principles of statics—equilibrium of forces and moments. However, it is unfortunately not entirely reliable in all aspects of further analysis, particularly in the interpretation of results.

## Artificial intelligence and GPT-4o

Artificial intelligence (AI) focuses on those aspects of human creativity that are related to intellectual work." Artificial intelligence (AI) is technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy", [1].

The development of artificial intelligence can be understood through a series of concepts that have emerged from one another over the past decades. This is illustrated in Figure 1.

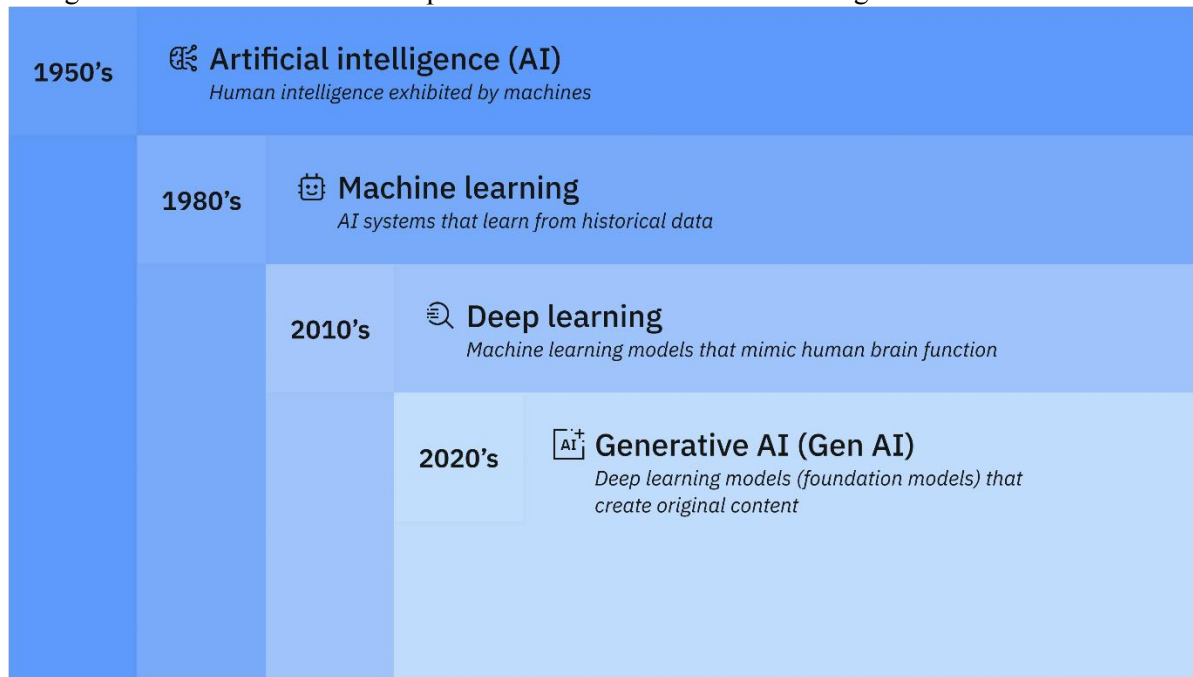


Figure 1: Series of nested or derivative concepts that have emerged over more than 70 years, [1]

Some of the advantages offered by AI include the automation of repetitive tasks, fast data processing, and decision-making with fewer human errors. We decided to use a model named GPT-4o because it allows for an image as input and produces text that presents formulas in a suitable LaTeX format. It represents an enhancement of the GPT-4 model. "GPT-4o ("o" for "omni") is our most advanced model. It is multimodal (accepting text or image inputs and outputting text), and it has the same high intelligence as GPT-4 Turbo but is much more efficient—it generates text 2x faster and is 50% cheaper." [2]

Research with models similar to ours exists, but only in recent times, as OpenAI launched ChatGPT 3.5 in late November 2022, which garnered significant attention from the general public. One of examination involves a manually crafted exam encompassing 126 multiple-choice questions, spanning various aspects of mechanics courses, including Fluid Mechanics, Mechanical Vibration, Engineering Statics and Dynamics, Mechanics of Materials, Theory of Elasticity, and Continuum Mechanics, [3]. This research has shown that LLM perform well when it comes to deductive reasoning, especially when the principles of prompt engineering are applied.

Some reviews focus on ChatGPT's performance on STEM-related exams and datasets. While methodologies, the version of ChatGPT used (i.e., GPT-4, GPT-3, ChatGPT, etc.), and the decision to use prompt engineering to improve performance vary among these studies, the results provide valuable insight on ChatGPT's strengths and weaknesses, [4]. The conclusion is that all models make arithmetic errors as well as unit errors, but they present these incorrect results with considerable confidence. It has been shown that anyone using ChatGPT to solve engineering problems must be prepared to verify the accuracy of the results. The advantage is that ChatGPT offers the ability to ask follow-up questions, although it should be emphasized that even using prompt engineering does not guarantee a correct answer.

Some authors have experience in connecting AI with other engineering fields. One example of this is work on tasks in the field of hydrodynamics [5]. Another application has been in the area of air quality management [6, 7].

## 2. Experiment with the application of GPT-4o in solving static beams

The idea was to investigate how GPT-4o can be applied to solving tasks in statics. To start, a simple example of a beam subjected to a vertical force, as shown in Figure 2. This simple task, formulated as follows: For the beam in the Figure 2, determine the maximum value of the bending moment and the location where it acts. The value of  $F = 2 \text{ kN}$ .

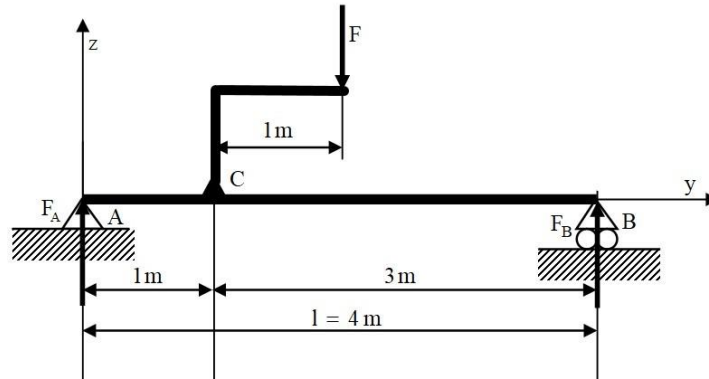


Figure 2: The example of a beam subjected to a vertical force, [8]

The task solution obtained through classical calculation:

$$\sum y_i = F_A - F + F_B = 0, \quad (1)$$

$$\sum M_A = F_B \cdot 4 - F \cdot 1 - M = 0, \quad (2)$$

$$F_B = \frac{F \cdot 1 + M}{4} = \frac{2 \cdot 1 + 2}{4} = 1 \text{ kN}, \quad (3)$$

$$F_A = F - F_B = 2 - 1 = 1 \text{ kN}, \quad (4)$$

The resulting diagrams of transverse forces and bending moments obtained from the classical calculations are presented in Figure 3. The value of the maximum bending moment, as  $M = 3 \text{ kNm}$ , acting at point C, is clearly indicated on diagram  $M_s$ . The maximum value of the bending moment was indeed the main question for GPT-4o.

GPT-4o outputs results in LaTeX format, which can be challenging to interpret. To facilitate a more accessible representation of mathematical expressions, a converter [9] can be employed to transform these formulas into a more readable format. For improved clarity, a tabular overview will be provided (Table 1), with the left column containing the original GPT-4o solution, and the right column displaying the corresponding expression processed through the converter. In the third column, it is indicated which answers GPT provided as correct and which ones were not.

GPT-4o provides the answer as following:

User: For the beam in the picture, determine the maximum value of the bending moment and the location where it acts. The value of  $F = 2 \text{ kN}$ . for uploaded image, on <https://i.imgur.com/L2fvLrW.jpeg>.

Assistant: To determine the maximum bending moment and its location for the given beam, follow these steps:

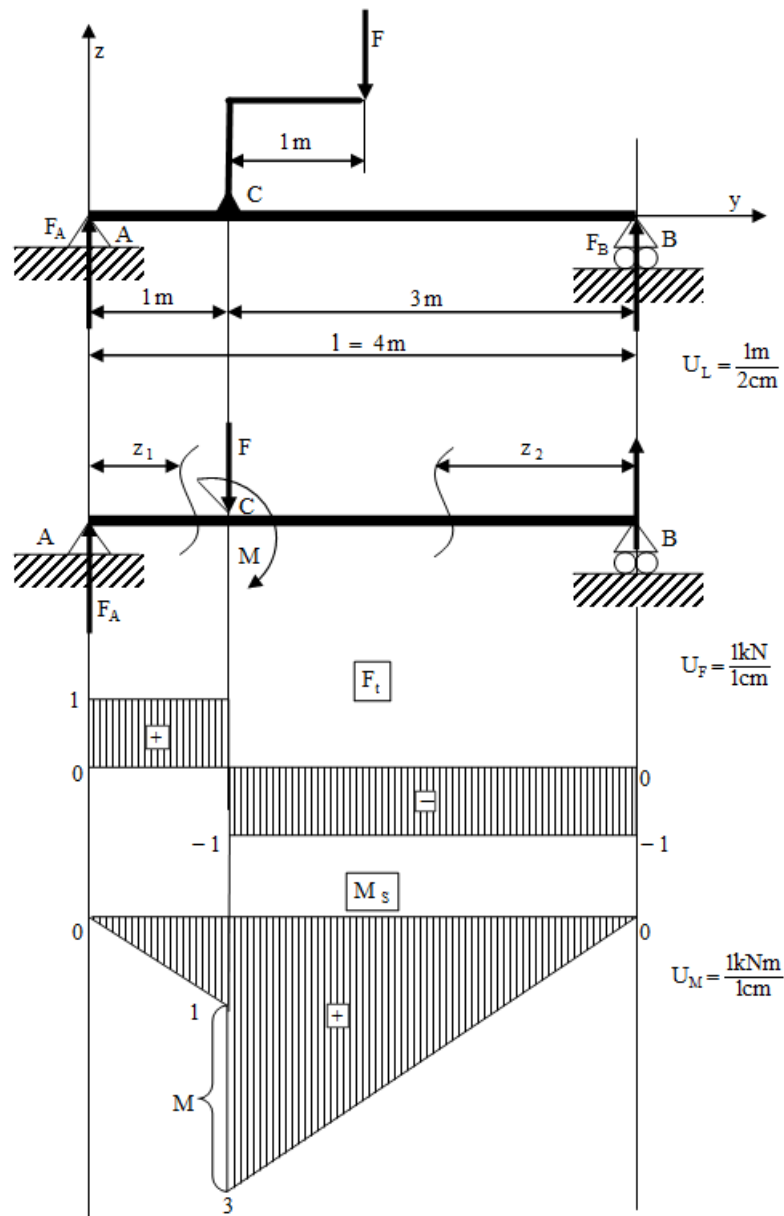


Figure 3: The resulting diagrams of transverse forces and bending moments, [8]

**Table 1:**

A tabular overview with the left column containing the original GPT-4o solution, the middle column displaying the corresponding expression processed through the converter and the right column shows answer status.

GPT-4o outputs	Corresponding expression processed through the converter	Answer status
### Step 1: Calculate the reactions at the supports	-	-
1. **Sum of vertical forces is zero:** $\sum F_y = 0$ $F_A + F_B - F = 0$	Sum of vertical forces is zero: $\sum F_y = 0 \quad F_A + F_B - F = 0$	correct
2. **Sum of moments about point A is zero:** $\sum M_A = 0$ $F_B \times 4 \text{ m} - F \times 2 \text{ m} = 0$	Sum of moments about point A is zero: $\sum M_A = 0$	correct

$\sum F_y = 0$ Since $(F = 2 \text{ kN})$ , $(4F_B - 4 = 0)$ $(F_B = 1 \text{ kN})$	Since $F = 2 \text{ kN}, 4F_B - 4 = 0$ $F_B = 1 \text{ kN}$	
3. <b>Using vertical force equilibrium:</b> $(F_A + F_B - 2 = 0)$ $(F_A + 1 - 2 = 0)$ $(F_A = 1 \text{ kN})$	Using vertical force equilibrium: $F_A + F_B - 2 = 0$ $F_A + 1 - 2 = 0$ $F_A = 1 \text{ kN}$	correct
### Step 2: Calculate the bending moment at key points.	-	-
#### Bending Moment at Point A: $(M_A = 0)$ (since the beam is pinned at A).	Bending Moment at Point A: $M_A = 0$ (since the beam is pinned at A).	correct
#### Bending Moment at Point C: Consider the beam section between A and C (left to right). $(M_B = F_A \times 4 \text{ m} - F \times 2 \text{ m})$	#### Bending Moment at Point C: Consider the beam section between A and C (left to right). $M_B = F_A \times 4 \text{ m} - F \times 2 \text{ m}$	correct
$(M_B = 1 \text{ kN} \times 4 \text{ m} - 2 \text{ kN} \times 2 \text{ m} = 4 \text{ kN} \cdot \text{m} - 4 \text{ kN} \cdot \text{m} = 0)$	$M_B = 1 \text{ kN} \times 4 - 2 \text{ kN} \times 2 \text{ m} = 4 \text{ kN} \cdot \text{m} - 4 \text{ kN} \cdot \text{m} = 0$	correct
### Step 3: Analyze intervals and draw the bending moment diagram.	-	-
- From A to C (0 m to 1 m): $(M(x) = 1 \text{ kN} \cdot x)$ This linearly increases from $(0 \text{ kN} \cdot \text{m})$ at $(x = 0)$ to $(1 \text{ kN} \cdot \text{m})$ at $(x = 1 \text{ m})$ .	- From A to C (0 m to 1 m):, $M(x) = 1 \text{ kN} \cdot x$ This linearly increases from $0 \text{ kN} \cdot \text{m}$ at $x = 0$ To $1 \text{ kN} \cdot \text{m}$ at $x = 1 \text{ m}$	correct
- From C to B (1 m to 4 m): $(M(x) = M_C = 1 \text{ kN} \cdot \text{m})$ The bending moment remains constant at $(1 \text{ kN} \cdot \text{m})$ in this interval	- From C to B (1 m to 4 m): $M(x) = M_C = 1 \text{ kN} \cdot \text{m}$ The bending moment remains constant at $1 \text{ kN} \cdot \text{m}$ in this interval	incorrect
### Step 4: Identify the maximum bending moment.	-	-
The maximum bending moment is $(1 \text{ kN} \cdot \text{m})$ , and it occurs at point C.	The maximum bending moment is $1 \text{ kN} \cdot \text{m}$ and it occurs at point C.	incorrect
### Conclusion:	-	-
<b>Maximum Bending Moment:</b> $(1 \text{ kN} \cdot \text{m})$	Maximum Bending Moment: $1 \text{ kN} \cdot \text{m}$	incorrect
<b>Location of Maximum Bending Moment:</b> $(1 \text{ m})$ from point A (at point C).	Location of Maximum Bending Moment: $(1 \text{ m})$ from point A (at point C).	correct



As we can see in the presented table, GPT-4o applies the correct equilibrium conditions and writes accurate equations for the balance of all transverse forces and bending moments. As a result, it produces correct support reactions. However, later on, when it needed to determine the maximum value of the bending moment and the location where it acts on the beam, it provided the correct answer for only one of the two; namely, that the maximum bending moment occurs at point C, while it made an error in its value.

### 3. Conclusion

The results obtained using GPT-4o show that when it comes to solving static beams, specifically determining characteristic static quantities in beam structures, we still cannot fully rely on the results. This holds true even for simple beam structures with basic loads, as demonstrated in the paper. Although GPT-4o provides the capability to solve a task based only on an uploaded image and the formulated question, which is a significant innovation, it is clear that this technology still needs further development to achieve greater reliability of results. On the other hand, we must emphasize that it is a major advancement that such a possibility exists to obtain a solution to a task simply from an uploaded image, opening entirely new perspectives. This indicates that development will continue toward further enhancement, and that these shortcomings will be addressed in the near future, representing a huge leap for the application of AI in the engineering world as well as in the educational process.

#### References:

- [1] What is artificial intelligence (AI)? URL: <https://www.ibm.com/topics/artificial-intelligence>.
- [2] GPT-4o. URL: <https://platform.openai.com/docs/models/gpt-4o>.
- [3] J. Tian, J. Hou, Z. Wu, P. Shu, Z. Liu, Y. Xiang, ... & X. Wang, Assessing Large Language Models in Mechanical Engineering Education: A Study on Mechanics-Focused Conceptual Understanding, 2024. arXiv preprint arXiv:2401.12983.
- [4] M. E. Frenkel, H. Emara, ChatGPT-3.5 and-4.0 and mechanical engineering: Examining performance on the FE mechanical engineering and undergraduate exams, Computer Applications in Engineering Education, e22781, 2023. <https://arxiv.org/pdf/2309.15866>.
- [5] Ž. Eremić, V. Mladenović, Primena ChatGPT kao podrška pri rešavanju elementarnih zadataka iz hidrodinamike, DIT Naučno stručno informativni časopis, volume 41, 2024, pp. 63-74.
- [6] D. Šešlija, Ž. Eremić, V. Mladenović, Istraživanje mogućnosti primene ChatGPT u upravljanju kvalitetom vazduha pod pritiskom, Naučno – stručni skup Preduzetništvo – inženjerstvo - menadžment PIM10, Zrenjanin, 8. jun 2024.
- [7] D. Šešlija, Ž. Eremić, V. Mladenović, V. Reljić, ChatGPT and compressed air quality in critical automated systems, 5th DQM International Conference, ICDQM-2024 Life Cycle Engineering and Management, Prijedor, Serbia, June 27-28, 2024.
- [8] C. Žepinić, Mehanika, Statika, Visoka tehnička škola strukovnih studija u Zrenjaninu, 2003.
- [9] Sciweavers. URL: <https://www.sciweavers.org/free-online-latex-equation-editor>.

# Six-Layered Application of Artificial Intelligence in Business Intelligence

Veselina Naneva<sup>1</sup>, Kremena Stefanova<sup>2</sup>

<sup>1, 2</sup> Paisii Hilendarski University of Plovdiv, 236 Bulgaria Blvd., 4003 Plovdiv, Bulgaria

*vnaneva@uni-plovdiv.bg; kstefanova@uni-plovdiv.bg*

## Abstract:

Business Intelligence (BI) represents the assemblage of several related stages, each of which works independently to process the required elements ensuring meaningful visual insights. Taking into account the specifics of each layer, we would consider a fully followable structure which can be applicable to every data-driven practices regardless of the business size or connection of the used tools. In addition to the described BI process essence, we will suggest enhancement steps in every level of it by the help Artificial Intelligence (AI). Furthermore, we will expand an already formed Five-layered architecture by adding an extra point, where the end user can actually participate in forming a new knowledge using the possibilities of Natural Language Processing (NLP) as subfield of AI.

## Keywords:

Business Intelligence, Artificial Intelligence, NLP application

## 1. Introduction

In recent decades businesses increasingly use a variety of software technologies for supporting their daily activities. These tools generate considerable amount of data which is coming from different sources, for instance operational procedures, finance transactions, customer interactions and others. As a result of this abundance, organizations are looking for decision-making strategies which are based simultaneously on past and current corporate information. In addition, they are trying to find the best solutions for administering these diverse types to provide some future insights and better understanding of the enterprise behavior. The high level of automation, connectivity and digitization produces enormous volumes of unstructured and structured data sets on a daily basis and it is impossible to obtain true value and sustainability without converting them into useful knowledge. It is essential to analyze and access such huge amounts of data to design certain plans, forecasting sales and demands, supply chain management, fast merchandising, and others, to compete with challenges in the business world [1, 2]. Precisely for this reason, the application of the Business Intelligence sphere is becoming more and more necessary. The term Business Intelligence (BI) refers to technologies, applications and practices for collection, integration, analysis, and visualization of business information across enterprises and organizations [3]. There are plenty of micro services that can be as well integrated in order to have an adequately driven BI algorithm flow.

To understand what the core structure of an enterprise Business Intelligence architecture is, we need to specify how it is divided by. In the scientific article titled “A Five-Layered Business Intelligence Architecture”, the BI process is split into five segments with actions that need to be taken into account regardless of the database size [4], i.e. the metadata storage layer, the resource’ one, the ETL (abbreviation of Extract, Transform, Load) steps, the stage of storing the information in an appropriate way and the last presentation one where the actual visual insights took place. Following their specifics and using the corresponding dependencies, the development of a meaningful data report will be guaranteed.

Although the abbreviation of AI also contains Intelligence, this topic is completely distinct from BI. Artificial Intelligence is a rapidly expanding field with the potential to revolutionize the way we live and work [5]. It relates to the ability of algorithms to clone human intelligence and perform tasks that are typically valid for human intelligence, such as learning, problem-solving, decision-making, and

natural language understanding. Data plays a crucial role in the concepts of the AI, as it provides the mentioned actions over time. Training an AI model demands a significant amount of information so the model can identify patterns, make accurate predictions, and continuously improve.

Using the key aspects of both spheres, modern business can receive full potential of predictive analytics which in real time can supply new knowledge. In order to define the actual working flow for the formed relation between the BI and AI, we will specify the necessity of acting in accordance with six-layered architecture for bringing a full capacity of data reporting for a certain enterprise.

## 2. Six-Layered BI Architecture

In this section we will suggest a six-layered BI architecture (SLA) and each of them will be covered in case of performing independently and simultaneously with the AI applied because of the business essence complexity and the respective approach.

### 2.1. Metadata Storage Layer

Metadata incorporates the information with more context or meaningful relevance. In consequence of it, we can define database sizes, data types, sources relations, duplications and other. In this layer we are operating on extremely unstructured level to process the semantic meaning of an individual piece of data and attributes that determine how the data is used or created [6]. This category controls naming, matching and parsing protocols, default values or their limitations. Semantic metadata assists tasks and rules for the collection and analysis of incoming data, which contribute structures and relations for the next phases.

The aim of metadata application can be divided into several contexts, such as [7]:

- **Supplementary context:** The metadata gives an explanation or the stored unstructured information. It depicts the source of the data's origin (where it came from) or its base format (how it is structured if such is recorded).
- **Logical structures:** Metadata streamlines how certain collection of data is organized and defines the connections between data elements, the existing dependencies and their hierarchy within the larger dataset. This narrative breakdown allows a logical and well-organized data structure, which makes the further actions more efficient.
- **Access control:** By providing information about user roles, data sensitivity levels, and other attributes relevant to access control, it is required to manage the access data permission although the information is not yet well structured.
- **Streamlined compliance:** The level of metadata can be depicted as a translator stage, where it can enable consistent data-handling practices across various systems. This can minimize errors and inconsistencies that could lead to compliance issues.

In addition to the corresponding context, this layer can be split into three segments which influence the correctness of the used tools, i.e. it can be specified as descriptive, structural and administrative. The first one encompasses essential features about the data resources taking into account its content, context, and characteristics. The second one addresses the hierarchical relationships between components, enabling productive data exploration and analysis. The last gives more details about maintaining data integrity and compliance across sectors.

Extracting metadata in a standardized way is effective for further engineering support for AI development [8] because it facilitates large-scale analysis and instruments in research and practice that manage multiple varying models. The existing AI concepts are diverse and that is why every AI application of this stage begins with establishing a metadata schema to outline these models and where they come from. As an example of the model metadata schema overview the impact of the formed model into layer essence is considered. The typical metrics which can be extracted here like name, tags, domains, references and so on are expanded by training algorithms. Following the concept, proposed by J. Tsay, A. Braz, M. Hirzel, et al, informative shards can be prepared for the stage of resource classification where if it is necessary a further action for their descriptive analysis will be done.

## 2.2. Resource Layer

The resource layer refers to descriptive analysis of the organizational data sources. For each company, there will be more than one type. Precisely for this reason, the level of the SLA has been separated and it illustrates the division of two key segments. Bear in mind the first one, the sources can be classified as primary data elements. They involve only raw and unstructured forms of information. Mainly this type requires additional steps for categorization, typification and normalization. After the base phase of the metadata has proceeded, for the stage of working with primary fields, the developer should use methodologies to create relationships. Through them, the sources can move to the operations for classification, becoming secondary data elements, while the other main one covers the already known base structure of the resources. At least the root environment can be specified. Since every business operates with a variety of tools, the topic can be split into two divisions: local and external. It is important to define where the data is coming from since for every secondary level section, individual approaches are essential.

As a result of AI applications, we can partially add three improvements. Using the machine learning algorithms, we can try to automate the discoveries. That approach will analyze many network resources with the same behavior and thus will identify new ones automatically. Furthermore, there are plenty of ready-based ML models for assessing their quality. They test the structure consistency of the database and validate every record origin individually from the global perspective. By these actions, the duplicates and anomalies can be found. Applying extra models for data training, significant patterns, trends and dependencies not visible in traditional analysis can be found in the stage of SLA.

## 2.3. ETL Process Layer

The definition of ETL covers operations for extracting, transforming and loading data which is firstly accessed in raw and unstructured type. We should specify individually the three mentioned topics, due to the fact that by them BI can support the various data sources which diversify and grow in volume.

The phase of extraction includes steps for retrieving information from disparate systems such as local files and external sources. This is often complicated not only by differences in formats, structures and access levels, but also by the need to be handled real-time data streams and services, or cloud-based platforms, which require scalable solutions for real-time extraction without manipulating the operational systems [9]. There are two types of extraction:

- full – all the fields from the root system are extracted, no matter of previous types of pulling out;
- incremental – the action is applicable only to the informative elements that have changed since last similar cycle.

In the transformation stage, the data should be normalized and reshaped into proper schemas by removing duplicated, managing missing records and ensuring its consistency. Such procedure is crucial for the further data analyzation. Except working with the low-level structures, transformation layer benefits finding an appropriate aggregation for summarizing detailed information into higher-level one. As a result, the performance is improved and advanced analytical tasks are enabled [10]. Usually, data from different resources must be combined during the transformation by matching records across many datasets based on common keys or attributes that are highlighted in the metadata phase or the resources one. Efficient joining algorithms and practices are demanded to manage performance and handle large volumes of data for the next stage of ETL methodology.

The load level involves placing the transformed information into a target system, typically a warehouse or directly in a visualization tool. The choice of a loading strategy, which can again be divided on incremental or fully up to date, depends on the size of the data and its complexity. Moreover, business requirements should be considered. Traditionally, the loading stage was designed for data collections methods, where large amounts of records are loaded periodically (daily, weekly, etc.). This approach is suitable for environments where real-time updates are not critical. For example, a case study can illustrate the monthly salaries in a company, paid on every last Friday of the month. However, in industries like e-commerce and finance, where there is a necessity of correction and analyzation in real-time dynamically, the loading mechanisms have become more prevalent.

With application of AI, the entire ETL process could undergo improvement and optimization. In each of the mentioned phases, the AI can take place together with the already formed algorithms. The extract and transform steps in ETL, particularly when dealing with unstructured and semi-structured data can be enhanced by AI techniques such as Natural Language Processing (NLP), working with ready-based models, and advanced parsing algorithms. As an example, we would like to suggest the Azure tools, part of the cloud service provider Microsoft Azure, which represent fully understandable user interface for operating with transformation and manipulation in order to find relations and classifications. User-unique models combine scenarios where the developer can easily identify the document type before invoking the respective extraction model [11]. They support many format types and can access fields from variety of documents, applying machine learning algorithms.

Technical sheets are generally represented as PDF documents and introduce product names, document titles and subtitles, other information about the products, property information in tabular format and the logo of the manufacturers, etc. For extracting text from similar PDF articles with complex structures, several open-source tools are available. There are plenty of custom developed models for table recognition, invoice components extraction and other which are also applicable to the AI usage subject matter.

## 2.4. Data Storage Layer

Data storage section is an essential part of the BI logic where the corresponding extracted, transformed and structured data can be appropriately stored. The three main types of data storage, namely Data Lake, Data Warehouse and Data Mart, contain distinct aspects.

Data Warehousing represents a critical architecture in data management perspective, serving as a bridge between the comprehensive data analysis and decision support systems. At its core, Data Warehousing involves the consolidation of data from various sources into a centralized repository, designed to facilitate efficient querying, reporting, and analysis [12]. Its architecture typically has three key components: the data fragments layer, the Data Warehouse layer, and the presentation layer. The first one consists of heterogeneous data resources as transactional databases, external data feeds, and legacy systems, which are integrated into the Data Warehouse through (ETL) operations, mentioned in the previous layer. The next one, also known as the data repository, is characterized by its data model, which may include star schemas, snowflake or galaxy schemas, intended to support efficient querying and analysis. Finally, the presentation layer contributes with tools and interfaces for data retrieval or working with the records selections.

The Data Lake's architecture, on the other hand, is planned to store large quantities of data in its raw form until it is needed for analysis. Unlike traditional Data Warehouses, which require data to be structured and transformed upon ingestion, Data Lakes allow the structure to be applied only when the data is queried or analyzed. This flexibility gives a significant advantage when dealing with heterogeneous or unstructured information shards. Data Lakes, especially those built on cloud infrastructure, support real-time analytics by enabling the storing and processing of streaming data. That capability is crucial for industries such as e-commerce, where businesses need to respond to constantly changing market conditions. In addition, conceptually Data Lake can be split in two key segments [13]:

- Graph-based models, which describe relationship between data and their metadata, illustrated in the first layer of the SLA;
- Data-Vault models, which are responsible for handling the complexity of accessed data, collected from different sources.

By enabling organizations to store data in its raw form, Data Lakes offer greater agility and scalability than traditional Data Warehouses, supporting advanced analytics and real-time insights. However, to fully realize their benefits, the corporate clients must address challenges related to data governance, security, and integration.

Data Mart, in turn, refers to a Data Warehouse subset, which is created to follow a specific business function or department requirements. In contrast to the ideology of Data Warehouse for storing the large amount of information, the Data Mart logic follows the centralized scope and there are two divisions with respect to the relationships – independent, if the segment of the Data Mart is unassociated with

any other for the enterprise, and dependent, if there is a recognizable connection with the rest structures. Regardless of the type, it improves business practice by accelerating query responses that work with smaller datasets and predict data trends [14]. Additionally, flexible multi-perspective decision-making is facilitated.

If we factor into the AI application, Data Lakes support Machine Learning (ML) operations as they provide the raw data necessary for building, training, and testing ML models. The AI-driven features prevent Data Lakes from turning into data swamps and enable more intelligent querying and analysis. The major impact on the data storage level based on AI application is the automation in AI model training and deployment within data warehousing systems, as it directly influences on the efficiency, scalability, and effectiveness of data analytics actions [15]. As an example for Microsoft Azure AI tools, we can consider the Automated Machine Learning (AutoML). This is a process of identifying classification, forecasting, regression tasks for the unstructured business data, specifying the source of the labeled training data and configuring the automated machine learning parameters [16], so that we can skip several steps of data analyzation while storing, which is illustrated on the Fig. 1. On the left side of it we can see three steps for initial processing of the structures, training the data model and data flow estimation action. Using the AutoML possibilities we can skip those three and combine them into a singular operation, automatically covering their behavior. The approach would bring the same result, but in an optimized and AI-driven way.

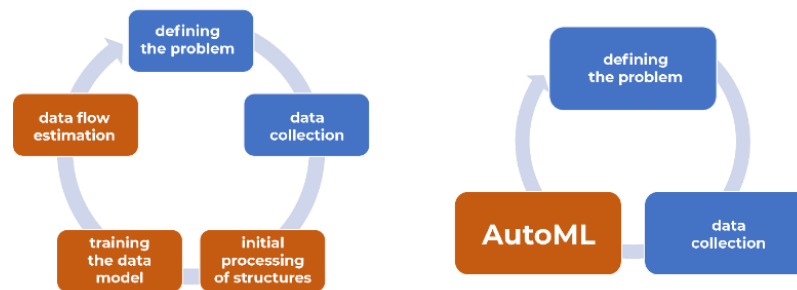


Figure 1: AutoML process improvement

## 2.5. Data Visualization Layer

One of the final layers in the SLA is the stage of data visualization. It defines the possibilities of a well-prepared and stored flow of information to be appropriately represented. The reason for dividing it into separate levels is the complexity of building the right report to answer the custom business questions. There are plenty of software products that satisfy all the necessities for optimized real-time accessed dashboard. Their core structure is split into three sections, such as data model, data presentation and report view. In short, they refer to:

- working with a structured database and manipulating the relationships of its tables, with reference to data model;
- transforming and managing the fields coming from the database, with regards to data presentation;
- adding a variety of informative visualization, synchronized with each other, in respect of the report view.

All of the BI end tools are following an equal algorithm. If we use a specified programming language, a field or combination of them from the database, we can add it into a calculation for forming a brand new knowledge. These elements can be included as a main definition of every visualization provided from a certain software product.

Of utmost importance is to clarify the visualization meaning. All of them provide a different type of purpose, accepted kind of data and context. The effectiveness of a dashboard lies upon its capacity to tailor the visual narrative to the corresponding requirements and objectives of its intended audience. By selectively filtering and emphasizing key data points, dashboards enable a more focused analysis that is aligned with organizational goals or individual project needs. The optimal choice of data visualization is determined by the nature of data and its presentation purposes, while most common of them are:

- Line Charts – track changes over time. Often the end user needs this type of element when he/she wants to see the actual rolling movement of a certain metric through time.
  - Bar Charts and Histograms – compare categories and distributions, which are suitable for group of elements that can be analyzed by their amounts in a visually appropriate way.
  - Pie Charts – focus on the percentages of each element segment from the 100% perspective.
- The selection of a certain visual provides the right data summarization and gives the ability to the end user to make informative decisions.

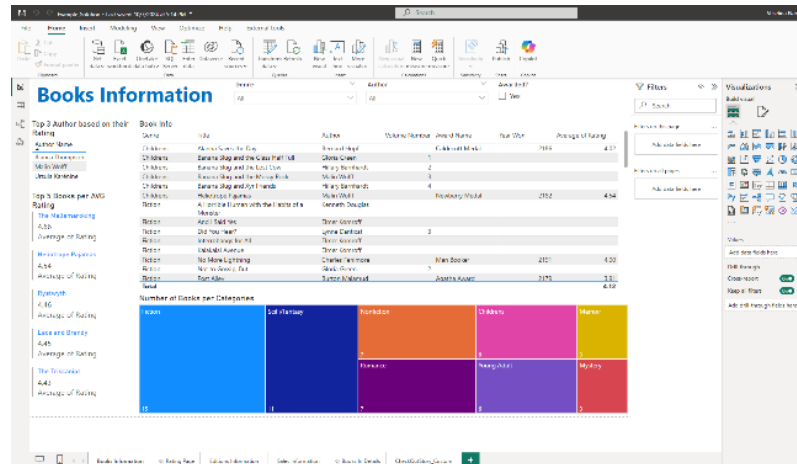


Figure 2: Report sample

Last but not least in this phase data security has been applied from an end user point of view. If we look for an example, the Microsoft BI tool Power BI supports a data restriction method called Row-Level Security (RLS), which works at row level, and in addition, filters within roles can be defined as well. In the Power BI service, privileged users to a workspace have access to semantic models in it. RLS only controls data access for users with Viewer permissions and provides limited data records based on the rules applied [16].

In current stage of SLA there is a semi-AI application in the perspective of Q&A visualization existence. The end user can retrieve the option to ask a case-formed inner interface in order to receive an informative answer.

## 2.6. AI Driven Layer

The final stage of the SLA represents AI as a participant in the operations of new knowledge generation. We can consider the impact of the natural language processing as a key approach for the layer. In short, NLP refers to the automatic ability of a human language to be recognized and analyzed. By the help of instruments for translation, pattern matching, sentiment analysis and voice recognition every well-written textual query can be modified into computer functions [17]. There are several NLP models which provide the appropriate result of the request given such as text classification, sentiment analysis, topic modeling and deep learning. Individually they are responsible for different types of data recognition so that the already provided human text or speech can be transformed into meaningful for the computer request. NLP methods are the core of the structure for the modern chatbots. By accessing the formed modes, they can handle standard tasks such as informing customers about products or services, answering their questions, etc.

Due to the fact that machine learning, as a subset of AI, works with a neural network of information from all available resources in the internet and existing relations with their metadata, the NLP application consumes huge computing power. The way human language has been interpreted requires multiple computer algorithm calculation actions which have an impact on the speed of receiving the necessary answer.

As a visual example of the NLP position in Business Intelligence, the Microsoft product Copilot will be specified. It provides deep insight into a certain company's data. In addition, it improves the simplifying business process and automation repetitive tasks [18]. The key advantages of its usage highlight three spheres of application, such as better understanding of the established visualization in

the report, developing brand new visual element found on textual demand for content and optimizing the calculation formula.

By dint of contextual description, the end user can easily receive a summarization of the global metrics in a report page, as it is presented in the Fig. 3. Although the information is not unknown for the responsible people, sometimes an additional inference should be made which can be difficult because of the only major indicators displayed. This problem can be solved by adding the Copilot solution into the environment. It receives the user natural language query, analyzes the already presented data and summarizes it in understandable way. From security perspective, the mentioned tool does not work beyond the RLS applied.

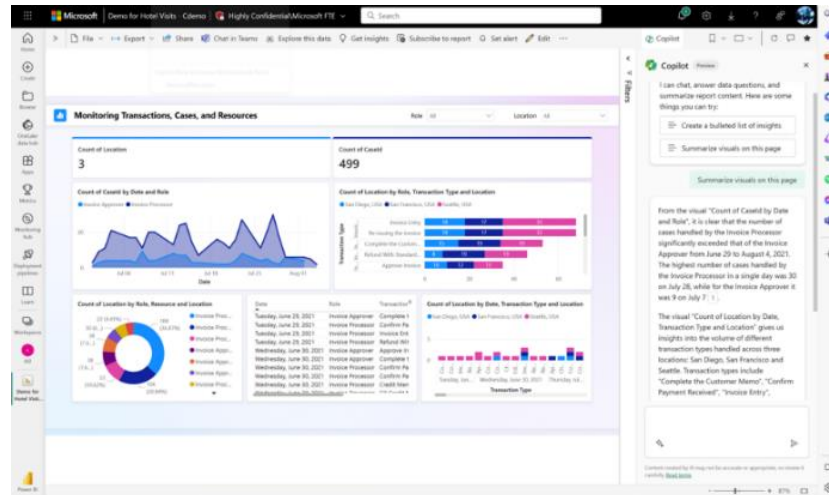


Figure 3: Copilot application

The second way of improvement by Copilot implies the real end user involvement in the visual generation. A dashboard comprehends limit number of representative elements as a primary structure. As we have already remarked, they are synchronized with each other. If the end user has own ideas or approaches for data analyzation, which are not in the report, Copilot gives an opportunity to be created a component, keeping the textual requirement. These new visualizations can illustrate data for the present or forecasting based on ML models which perform with the corresponding informative fields. An additional visualization does not enrich the report visual content and it is visible only to its author if certain option is selected.

The third AI-driven approach for using Copilot is the calculation formula enhancement. Through adequate user textual query, the tool can generate a fully functional and adaptive syntax for building new elements that can be used in the actual structure. The more optimized the computation is, the better dashboard performance will be.

### 3. Conclusions

Business Intelligence nowadays is widely applicable for analyzing of every enterprise work process. Regardless of the database complexity or company volume, we suggest a BI architecture that contains division of six individual layers, such as metadata stage, resource level, ETL operations, storage activities, data visualization flow and AI-driven phase, where the end user participate in accessing new knowledge through his/her own data by the help of NLP abilities. Although in the past the first five layers had been functioned singularly, in the era of the digitalization AI finds its place in all of the six levels described. Following the considered structure will provide well designed, real-time accessed data report which includes meaningful business inferences, dynamically managed by the responsible user.

### Acknowledgment:

This paper is partially supported by project MUPD23-FMI-009 of the Scientific Research Fund of Paisii Hilendarski University of Plovdiv, Bulgaria.



## References:

- [1] F. Bordeleau, E. Mosconi, L. Santa-Eulalia, Business intelligence in industry 4.0: State of the art and research opportunities, *Proc. of the 51<sup>st</sup> Hawaii International Conference on System Sciences*, Waikoloa Village, HI, USA, 2–6 January 2018.
- [2] S. Ahmad, S. Miskon, R. Alabdan, I. Tlili, Towards Sustainable Textile and Apparel Industry: Exploring the Role of Business Intelligence Systems in the Era of Industry 4.0, *Sustainability*, 12 (7), 2632, 2020, <https://doi.org/10.3390/su12072632>.
- [3] B. Zohuri, M. Moghaddam, From Business Intelligence to Artificial Intelligence, *Mod App Matrl Sci* 2 (3), 2020, MAMS. MS.ID.000137, DOI: 10.32474/MAMS.2020.02.000137.
- [4] I. Ong, P. Siew, S. Wong, A Five-Layered Business Intelligence Architecture, IBIMA Publishing, *Communications of the IBIMA*, Vol. 2011, 2011, Article ID 695619, 11 pages, DOI: 10.5171/2011.695619.
- [5] A. Aldoseri, K. Al-Khalifa, A. Hamouda, Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges, *Applied Sciences*, 13 (12): 7082, 2023, <https://doi.org/10.3390/app13127082>.
- [6] <https://www.redpointglobal.com/blog/what-to-know-about-metadata-at-the-data-layer/>
- [7] <https://www.astera.com/type/blog/metadata/>
- [8] J. Tsay, A. Braz, M. Hirzel, et al., Extracting enhanced artificial intelligence model metadata from software repositories, *Empir Software Eng*, Vol. 27, art. number 176, 2022, <https://doi.org/10.1007/s10664-022-10206-6>.
- [9] J. Wang, W. Yang, Y. Wang, R. Sherratt, J. Zhang, Big Data Service Architecture: A Survey, *Journal of Internet Technology*, Vol. 21, No. 2, 2020, pp. 393–405, DOI: 10.3966/160792642020032102008
- [10] <https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/concept-custom?view=doc-intel-4.0.0>
- [11] <https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/studio-overview?view=doc-intel-4.0.0&tabs=di-studio>
- [12] A. Ambasht, Real-Time Data Integration and Analytics: Empowering Data-Driven Decision Making, *International Journal of Computer Trends and Technology*, Vol. 71, Issue 7, 8-14, July 2023, pp. 8–14, ISSN: 2231–2803, <https://doi.org/10.14445/22312803/IJCTT-V71I7P102>
- [13] P. Vaish, S. Shricastava, S. Sen, Business Intelligence: Escalation of Data Warehousing and Data Mining for effective Decision Making, *International Journal of Advanced Science and Technology*, Vol. 29, No. 5s, 2020, pp. 1377-1388, ISSN: 2005-4238.
- [14] S. Rachakatla, P. Ravichandran, J. Machireddy, Scalable Machine Learning Workflows in Data Warehousing: Automating Model Training and Deployment with AI, *Australian Journal of Machine Learning Research & Applications*, Vol. 2, Issue 2, 2022, pp. 262–286, <https://sydneyacademics.com/index.php/ajmlra/article/view/109/104>
- [15] <https://learn.microsoft.com/en-us/azure/machine-learning/concept-automated-ml?view=azureml-api-2>
- [16] <https://learn.microsoft.com/en-us/fabric/security/service-admin-row-level-security>
- [17] Y. Kang, Z. Cai, C. Tan, Q. Huang, H. Liu, Natural Language Processing (NLP) in Management Research: A Literature Review, *Journal of Management Analytics*, Vol. 7, Issue 2, 2020, pp. 139–172, <https://doi.org/10.1080/23270012.2020.1756939>
- [18] <https://learn.microsoft.com/en-us/power-bi/create-reports/copilot-introduction#copilot-for-report-authors>

# Personal data protection in eGovernment

Rade Dragović<sup>1</sup>, Nadežda Ljubojev<sup>2</sup>, Dalibor Dobrilović<sup>2</sup>, Bojan Perović<sup>1</sup>, Milenko Đuričić<sup>3</sup>

<sup>1</sup> Institute for standards and technology, Belgrade, Serbia

<sup>2</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

<sup>3</sup> Misdemeanor Appellate Court, Belgrade, Serbia

*rade.dragovic@instate.biz; nadezda.ljubojev@tfzr.rs; dalibor.dobrilovic@tfzr.rs;  
bojan.perovic@instate.biz; milenko.djuricic@pkap.sud.rs*

## Abstract:

The introduction of information technologies into modern living environment has a significant impact on traditional social concepts and patterns of behavior both in private life and in the state administration system. The modern world cannot be imagined without new technologies, since information trends have influenced to all areas of social life in domain of the application of Internet and electronic services in all areas of life. The essence of the modern information society is reflected precisely through the exchange and use of a large amount of information by Internet and information and communication technologies in everyday life and business activities. The use of modern technologies defines some new technological doctrines such as information security and protection of personal data which need to be legally defined and introduced into business system of state authorities. In accordance with the importance of modern phenomena and the increase of their role in real life the main subject of our research is the legal aspects of personal data protection in electronic public administration. Using the example of the Serbian legislature, the paper will present the relationship between laws, by-laws and standards regulated by the domain of electronic public administration.

## Keywords:

data protection, personal data, information security, electronic government, law, legal regulative.

## 1. Introduction

Public administration has undoubtedly come into contact with new technologies and this contact has fundamentally influenced the legal nature of public administration, its relationship with citizens and its role in modern society. Thanks to modern technologies, public administration changes the nature of business and becomes electronic public administration with numerous specificities that this change brings as innovative. In the new technological, organizational and certainly procedural environment, the state administration performs tasks assigned to that state body more quickly and efficiently, achieves better communication with other legal and citizens and its work becomes completely transparent. There are numerous examples in most European countries that it not possible to communicate with the public administration in traditional paper form, but communication is done exclusively in electronic form (Tax administration, Business registers agencies, Geodetics authorities). Through the application of new technologies, citizens get an open, transparent, digital administration with the possibility of insight into the course of processing of administrative bodies' cases and easier access to services through frequent downloads of final documents without the traditional and with an electronic seal. In the new digital environment, state administration and citizens exchange large amounts of information/data. This generation of a large amount of data becomes the basis for state authorities for the development of machine learning and numerous elements of the application of artificial intelligence oriented towards the provision of better and faster services for legal entities and citizens. New phenomena also bring with them risks for the functioning of public administration and society. The use of information and communication technologies opens up important questions related to the legal and proper performance of administrative tasks. By

increase in the importance of information and data, there are also challenges for security, privacy and other rights of citizens. Therefore, new concepts and challenges require adequate and modern regulation based on theoretically analyzed points of view followed by practical analysis of such phenomena. [1].

## 2. Electronic government

Two reasons are crucial for the indispensability of e-government for any nation, regardless of its development and economic status. First, the use of technological initiatives by governments will contribute to their efficiency and competitiveness in the current environment. Second, e-government has the potential of enabling democratic governance, promoting democratic practices, and facilitating efficient contact between governments and the citizens. Electronic government has been defined as “the continuous optimization of service delivery, constituency participation, and governance by transforming internal and external relationships through technology, the Internet, and new media” (Gartner Group, May 2000). Thus, e-government can be considered through two lenses: the type of relationship and the stage of integration. E-government can involve electronic relationships between the government and different levels of constituents. Building on the categories suggested previously by other writers, we offer a more complete view of the multidimensional relationships between governments and the entities with which they interact. Types of E-government are [2]:

- **Government Delivering Services to Individuals (G2IS):** In this case the government establishes or maintains a direct relationship with citizens in order to deliver a service or benefit. This would include the Social Security Administration in its delivery of benefits, for example. It can also involve two-way communications as, for example, when individuals request information about benefits or the government needs information to process benefits.
- **Government to Individuals as Part of the Political Process (G2IP):** This is the relationship between the government and its citizens as part of the democratic process. It is perhaps the most essential relationship between a government and any entity. Examples include voting online and participating in requests for comments online during the regulatory process.
- **Government to Business as a Citizen (G2BC):** Although businesses do not vote, and thus the relationship between businesses and the government will not look exactly like the G2IP, there are still opportunities for businesses to relate to the government in a citizen-like capacity. Providing Securities and Exchange Commission filings online and paying taxes online would be examples of the relationship between government and businesses in this category.
- **Government to Business in the Marketplace (G2BMKT):** Companies receive a number of online services from government institutions in order to run their business as efficiently as possible. Efficiency was achieved by reducing paperwork, sending/receiving mail and time intervals waiting for deadlines for appeals to expire, ... State agencies group their work (most often by department, but there are good examples of grouping in an effort to make certain strategic projects easier for companies - for example, issuing construction permits) in order to support companies in doing better business and paying higher taxes to the state. These platforms can be used in other types of eGovernment such as local self-government bodies, which would lead to significant savings.
- **Government to Employees (G2E):** Online relationships between government agencies and their employees face the same requirements as the relationships between businesses and their employees. For example, government agencies can use an intranet to provide information to their employees and can typically allow some online transactions with their employees if they have the proper technological architectures. This relationship should be distinguished from the same individual's relationship under G2IP and G2IS.
- **Government to Government (G2G):** Government agencies must often collaborate and/or provide services to one another. There are substantial gains from conducting some of these transactions online. Government-to-government applications can be performed between federal agencies, or between federal, state and local agencies. An example of an inter-governmental level e-government application is the National Science Foundation's request that all proposals for

research funding by public academic institutions be submitted by an online application. The potential for G2G to benefit the government agencies involved is tremendous.

Electronic government encompasses all government roles and activities, shaped by information and communications technologies (ICTs). Going well beyond analogies to e-commerce, it encompasses the four domains of governance and public administration: the state's economic and social programs; its relationships with the citizen and the rule of law (e-democracy), its internal operations and its relationship with the international environment. E-government builds on three evolving forces: technology, management concepts and government itself. It has given rise to several phenomena that are redefining the public sector environment, including the International Institute of Administrative Sciences. Four aspects of e-government have lasting impacts on public administration: citizen-centered service, information as a public resource, new skills and working relationships, and accountability and management models. Public administration in all countries requires new thinking and leadership to ensure that e-government realizes its full potential [3].

From another point of view, eGovernment concept is initiated by the application of information technologies, but it is based on the following overlapping values:

- Improved public services;
- improved administrative efficiency;
- Open Government (OG) capabilities;
- improved ethical behaviour and professionalism;
- improved trust and confidence in government; and
- improved social value and well-being.

These six public value dimensions were thereafter generalized into three overarching and also overlapping, public value dimensions of Improved Public Services, Improved Administration, and Improved Social Value[4].

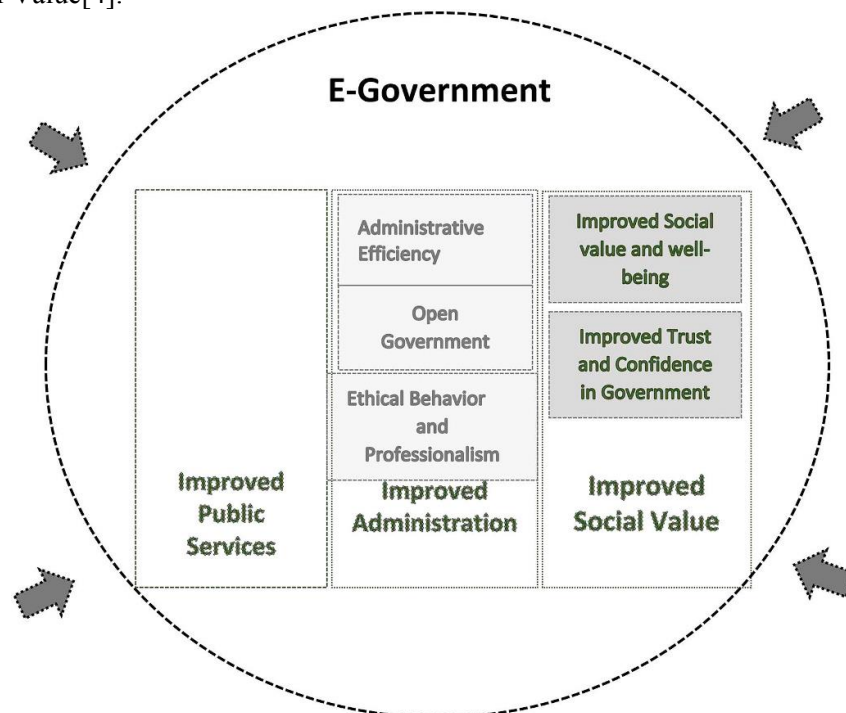


Figure 1: Generalization of the six dimensions of the public value of e-government[4]

### 3. Electronic Government regulative

The Law on Electronic Government defines the terms protection of personal data and information security, which makes eGovernment the based on the Law on Protection of Personal Data and the Law on Information Security, which in the domain of international and national standards are followed by standards ISO/IEC 27001 - Information security, cybersecurity and privacy protection -

Information security management systems - Requirements, ISO/IEC 27002 - Information security, cybersecurity and privacy protection - Information security controls and ISO/IEC 27701 - Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management - Requirements and guidelines, ISO 20000-1 - Information technology - Service management.

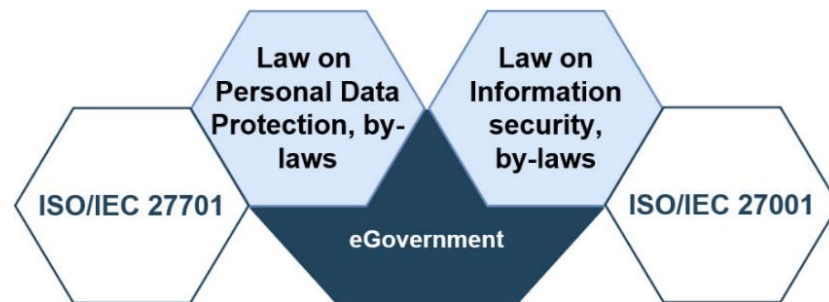


Figure 2: Relations of laws and standards according to eGovernment regarding privacy

The Laws related to the Law on Electronic Administration of the Republic of Serbia are: Law on Free Access to Information of Public Importance, Law on electronic media, Law on Public Media Services, Law on Electronic Document, Electronic Identification and Trust Services in Electronic Business, Law on data confidentiality, Law on special powers for effective protection of intellectual property rights, Law on Electronic Communications, Law on Information Security, Law on Payment Services and Law on the Protection of Personal Data, Law on inspection supervision, Law on tax procedure and tax administration, Law on tax procedure and tax administration, Law on the procedure for registration in the real estate cadastre and infrastructure cadastre, Law on the Central Population Register, Law on enforcement and security, Law on culture, Law on Digital Property, Law on social card, Law on the use of renewable energy sources, Law on energy efficiency and rational use of energy, Law on Energy, Law on mining and geological research, Law on the seal of state and other authorities, Law on electronic document, electronic identification and trusted services in electronic business, Law on Business Companies, Law on Agriculture and Rural Development, Law on Patents, Law on the registration procedure in the Agency for Economic Registers, Law on Occupational Safety and Health, Law on Electronic Communications.

The Strategies related to the Law on Electronic Administration of the Republic of Serbia in 2024 are: Personal data protection strategy, Strategy for the development of artificial intelligence in the Republic of Serbia, Public administration reform strategy in the Republic of Serbia, Strategy for the development of information society and information security in the Republic of Serbia, Digital skills development strategy in the Republic of Serbia, Strategy of sustainable urban development of the Republic of Serbia, Strategy for improving the position of persons with disabilities in the Republic of Serbia, Strategy of prevention and protection against discrimination and others.

There are a large number of strategies, decrees and regulations that are on legal base linked to the Law on the area of Electronic Administration.

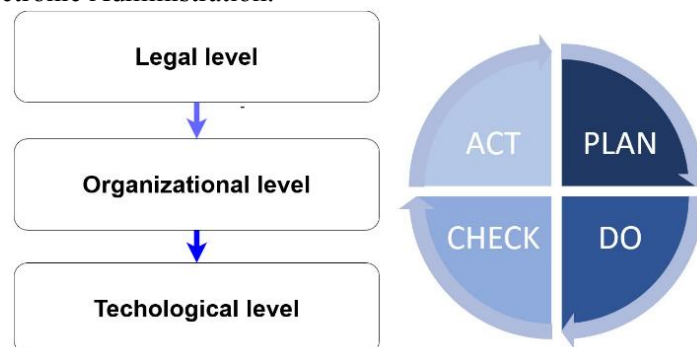


Figure 3: Law, technology and organizationrelations

The concept of the law that will regulate electronic administration also required by-laws accompanying this law: Regulation on the maintenance and improvement of the State Center for Data

Management state bodies web presentation and Storage, Regulation on detailed conditions for the creation and maintenance of the, Regulation on the manner of operation of the Open Data Portal, Regulation on the way of keeping the Meta data register, the way of approving, suspending and canceling access to the service bus of the body and the way of working on the eGovernment Portal, Regulation on organizational and technical standards for the maintenance and improvement of the Unified Information and Communication Network of electronic administration and the connection of authorities to that network, Regulation on further conditions for the establishment of electronic administration, Rulebook on the manner in which authorities inspect, obtain, process and transfer, i.e. deliver data on facts that are kept in official records from registers in electronic form which are necessary for decision-making in administrative proceedings. The mentioned framework of by-laws is not enough to fully define the technological framework that is constantly changing and improving.

It was observed that the Law on Electronic Administration is connected to all departments of state bodies, as well as that, apart from the legal domain, it is connected to the technological and organizational domain through special laws. Law on Electronic Administration of the Republic of Serbia is an example of a law that provides synergy for the functioning of information systems, harmonizing the technological domain, which is the initiator of better efficiency and effectiveness the state bodies with its organizational discipline and needs. In the practical application of the Law on Electronic Administration, it is necessary to continuously review the real basis and functioning of information systems based on the Deming cycle: Plan–Do–Check–Act cycle. From everything initially analysed and previously partially stated, it is clear that the Law on Electronic Administration is the paramount law for the functioning of the entire eGovernment System of a modern democratic state based on the rule of the law. The work of state bodies must be defined and regulated, but this is in the field of application of the ever-changing field of information technologies that are operationally and tactically changed through innovative solutions; it is simply not possible to fully define them in the field of legal and by-laws. For the missing part in the technological definition of information systems with numerous definitions of the organizational and procedural domain, advanced countries rely on ISO standards that have been adopted as international as well as national standards. In the domain of law, they often have the dimension of a by-law or rules of profession and science in the legal classification of judicial practice.

#### 4. Data privacy

The reasons for protecting privacy in public are the same as the reasons for protecting privacy in the earlier traditional environment. The enormous availability of data leads to the danger that certain data will be misused, especially through the application of the latest technologies that have not yet been fully defined by legislation, such as video surveillance or artificial intelligence. [5].

Personal data represent a characteristic feature of a specific person. They are used in everyday life situations, serving as a means of concretizing and realizing legal relations. However, personal data is not only a means of identification. In fact, they represent the appearance of a certain personality in legal life, which serve to realize various rights and interests.

One of the first determinations of personal data in comparative law can be found in the Convention on the Protection of Individuals in Relation to Automatic Data Processing, which was adopted by the Council of Europe in 1981. This convention provides that personal data is any information relating to an identified or identifiable person. In relation to other domestic and foreign regulations, this is the broadest definition. Since it was adopted more than 35 years ago, this provision represented a pioneering definition of personal data in the legal world, which has not changed much in the years that followed. We can say that the mentioned definition was used, in the amended and supplemented edition, in later European regulations.

The EU directive on the protection of individuals with regard to the processing of personal data and the free flow of such data from 1995 contained a definition of the concept of personal data. Within the framework of the Directive, personal data refers to any information relating to a specific or identifiable person, whereby the identifiable person can be identified directly or indirectly, in particular by means of an identification number or one or more specific factors related to his physical, psychological, economic, cultural or social identity.

The basis of this definition was followed by the most important European document in relation to data protection, the Regulation on the Protection of Citizens in relation to the processing of personal data and free movement of such data (EU General Data Protection Regulation), better known as GDPR (abbreviated from General Data Protection Regulation) with the entry into force of which the Directive from 1995 was repealed. The general EU regulation defines the concept of personal data by stating that it is personal data "all data related to a natural person whose identity is determined or can be determined; an identifiable natural person is a person who can be identified directly or indirectly, in particular by means of identifiers such as name, identification number, location data, network identifier or by means of one or more factors inherent to physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". In order for a piece of data to be defined as personal data, it must carry a characteristic that refers to a specific natural person/citizen or a citizen that can be determined, whereby the information obtained refers to a specific personal and unique property of a specific citizen, which is accepted solution from the Directive. Some of the personal data that includes the previous definition are: first name, last name, nickname, email address, address and postal code, business address or location, citizenship, data from birth register sources, passports, identity cards or travel documents, physical features (hair color, eye color), physiological features (data from health records), cultural identity (membership in cultural organizations), social identity (social media profiles), social security number, audio recordings, video recordings, photographs, phone numbers, data o location, site search history, digital signature, etc. The definition of the EU General Regulation is identical to the definition of personal data from the Personal Data Protection Act of Serbia [6].

In the area of personal data protection, special attention should be paid to the Law on Electronic Administration, which applies to activities in the domain: „Establishing and maintaining registers and records in electronic form, Protection of data and documents during their acquisition and transmission, Establishing and maintaining a register Meta data, Software solution, Payment on the eGovernment Portal, Conditions for establishing electronic administration, Conditions for obtaining and providing data and electronic documents and Certificate of electronic delivery refer to the Personal Data Protection Act“. The procedural framework for defining and implementing data privacy in countries of European provenance is the corresponding Law on the Protection of Personal Data (in European countries adopted on the model of the General Data Protection Regulation (GDPR) from 2016) and the international and national standard ISO/IEC 27701 Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management - Requirements and guidelines. The Law on the Protection of Personal Data protects all personal data except for the data of competent authorities for the purposes of preventing, investigating and detecting criminal offenses, prosecuting perpetrators of criminal offenses or enforcing criminal sanctions, including prevention and protection against threats to public and national security, as well as the free flow of such data.

There is a growing understanding that innovation, creativity and competitiveness must be approached from a “design-thinking” perspective. Privacy must be approached from the same design-thinking perspective. Privacy must be incorporated into networked data systems and technologies, by default. Privacy must become integral to organizational priorities, project objectives, design processes and planning operations. Privacy must be embedded into every standard, protocol and process. The 7 Foundational Principles of Privacy by Design are [7]:

1. **Proactive not Reactive.** Preventative not Remedial. The Privacy by Design approach is characterized by proactive rather than reactive measures. It anticipates and prevents privacy invasive events before they happen. Privacy by Design does not wait for privacy risks to materialize, nor does it offer remedies for resolving privacy infractions once they have occurred - it aims to prevent them from occurring. In short, Privacy by Design comes before-the-fact, not after.
2. **Privacy as the Default.** We can all be certain of one thing - the default rules! Privacy by Design seeks to deliver the maximum degree of privacy by ensuring that personal data are automatically protected in any given IT system or business practice. If an individual does nothing, their privacy still remains intact. No action is required on the part of the individual to protect their privacy - it is built into the system, by default.
3. **Privacy Embedded into Design.** Privacy by Design is embedded into the design and architecture of IT systems and business practices. It is not bolted on as an add-on, after the fact.

The result is that privacy becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality.

**4. Full Functionality - Positive-Sum, not Zero-Sum.** Privacy by Design seeks to accommodate all legitimate interests and objectives in a positive-sum “win-win” manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. Privacy by Design avoids the pretense of false dichotomies, such as privacy vs. security, demonstrating that it is possible to have both.

**5. End-to-End Lifecycle Protection.** Privacy by Design, have been embedded into the system prior to the first element of information being collected, extends securely throughout the entire lifecycle of the data involved, from start to finish. This ensures that at the end of the process, all data are securely destroyed, in a appropriate period. Privacy by Design ensures lifecycle management of information, end-to-end approach.

**6. Visibility and Transparency.** Privacy by Design seeks to assure all stakeholders that whatever the business practice or technology involved, it is in fact, operating according to the stated promises and objectives, subject to independent verification. Its component parts and operations remain visible and transparent to users and providers alike. Remember, trust but verify.

**7. Respect for User Privacy.** Above all, Privacy by Design requires architects and operators to keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options. Keep it user-centric.

## 5. Conclusions

The intersection of privacy interests and the implementation of information technology and electronic government to enhance efficiency and ease of use for citizens is not a simple topic, but a dynamic and multifaceted one. Paradoxically, but understandably, laws and executive orders both mandate action and restrict the government in its pursuit of these goals.

The development of information and communication technologies of state bodies has significantly relied on the resource of personal data used in the processing of requests, data exchange and storage in numerous dedicated databases. The new technological state circulation generates a large number of personal data for which it is necessary to define a clear legal basis that will be implemented as a technological requirement in the new software solutions of state authorities. There are numerous obstacles to a clear definition of the legal framework where numerous software initiatives are no longer limited to countries but operate globally.

Today, databases are usually stored in electronic form, stored in server/storage memory or in virtual form in "clouds". They represent a unique place where stored data of a certain category or type can be accessed, for their download and further use for specific purposes. For this reason, databases have found their place in legal systems, especially in relation to the activities of electronic public administration, which need to be organized in a technologically functional but also legally valid manner.

Nevertheless, the use of information and communication technologies may raise data protection concerns, for instance if an integral part of the public administration organization has a policy of open data that could compromise the privacy of some individuals or internal government data sharing among government institutions.

The use of public information and database processing in Serbian public sector helps efficiency and good governance, but should not neglect need for constant improvement the activities of implementation in terms of administrating public data and exchanging information with citizens and public servants in adherence to human rights.

A comprehensive legislative framework that guarantees human rights in e-governance and enhances the transparency, accountability, and responsiveness of public institutions, and is in line with the EU legal framework is a prerequisite for the development of eGovernment

The use of highest information and communication technologies security must standards must be implemented when public services are delivered through digital channels. Providing the reliable eGovernment can be achieved only by accordance with the international generally accepted standards for protection of citizen's privacy and personal data, with the provision of regulatory environment with the rising attention paid to cybersecurity of eGovernment.



### References:

- [1] Andonović, Stefan H. Data Protection in Electronic Public Administration in the Republic of Serbia: Legal Aspects. Diss. University of Belgrade (Serbia), 2019.
- [2] Hiller, Janine S., and France Bélanger. "Privacy strategies for electronic government." *E-government* 200.2001 (2001): 162-198.
- [3] Brown, D. (2005). Electronic government and public administration. *International Review of Administrative Sciences*, 71(2), 241-254. <https://doi.org/10.1177/0020852305053883>
- [4] Twizeyimana, Jean Damascene, and Annika Andersson. "The public value of E-Government—A literature review." *Government information quarterly* 36.2 (2019): 167-178.
- [5] Nissenbaum, Helen. "Protecting privacy in an information age: The problem of privacy in public." *The ethics of information technologies*. Routledge, 2020. 141-178.
- [6] Andonović, Stefan, and Dragan Prlja. "Osnovi prava zaštite podataka o ličnosti." (2020).
- [7] Cavoukian, Ann. "Privacy by design: The seven foundational principles." IAPP Resource Center, <https://iapp.org/resources/article/privacy-by-design-the-7-foundational-principles> (2021)
- [8] Dragović, R., Stanisavljev, S., Dobrilovic, D., Dragović, D., Milosev, V., Software information security management for government authorities, IIZS 2023, ISBN 978-86-7672-368-3

# Encryption as an information security function in a critical infrastructure systems

Vladimir Kačanovski<sup>1</sup>, Rade Dragović<sup>2</sup>, Dragan Dragović<sup>3</sup>, Bojan Perović<sup>2</sup>

<sup>1</sup> Basic court in Valjevo, Valjevo, Serbia

<sup>2</sup> Institute for standards and technology, Belgrade, Serbia

<sup>3</sup> Serbian Business Registers Agency, Belgrade, Serbia

vladimir.kacanovski@va.os.sud.rs;  
bojan.perovic@instate.biz

rade.dragovic@instate.biz;

ddragovic@apr.gov.rs;

## Abstract:

In the modern technological age, information security has assumed a very important technological and organizational role, which, through very precise requirements, lays the foundation for the safe functioning of information systems in the field of critical infrastructure. By applying encryption, potential risks are identified, treated and reduced/eliminated to the lowest possible, acceptable level. The paper aims to demonstrate the importance of encryption protection and information security in a critical infrastructure system. The primary purpose of encryption is to protect the confidentiality of stored and/or transmitted and downloaded data. Encryption techniques are applied to data to obtain encrypted data, this process is known as encryption. Information security and crypto-protection occupy a special place in the field of design, implementation and exploitation of critical infrastructure information systems and their protection. The paper will present guidelines for solving information security problems in critical infrastructure information systems in the field of encryption.

## Keywords:

information security, encryption, standards, data, protection, critical infrastructure

## 1. Introduction

Cryptography is a scientific discipline that deals with the study of methods for securely sending messages in a form that is understandable only to those for whom the information is intended - the end users, while the message will be unusable for everyone else. The parties participating in the communication through the communication channel are called the sender and the receiver. The message that the sender wants to send to the receiver is called plain text. The sender transforms the plaintext using a pre-agreed key in a process called encryption to obtain an encrypted message (code) to send over the communication channel. The receiver knows the key and can use it to determine the plaintext. This process of returning a message from an encrypted form to its original form is called decryption. A person listening to a communication channel is called an adversary, an eavesdropper, or an attacker. Attacks are malicious actions performed by unauthorized members of the communication model. A successful attack on a cryptographic system involves finding a practical way for an adversary to obtain the plaintext from the ciphertext. Informally, we'll often refer to this as a code violation. A cryptographic algorithm or cipher is a function used for encryption and decryption. They are actually two functions that map basic elements of plain text to basic elements of cipher, and vice versa. The mentioned functions are selected from a family of functions, depending on the key used. We call the common set of all possible key values the key space. A cryptosystem consists of a cryptographic algorithm and all possible plaintexts, ciphers and keys.

The aim of this paper is to provide basic guidelines on how to better understand the importance of encryption as an important tool in protecting confidential and sensitive information in a critical infrastructure system with concrete experiential recommendations in the domain of recommended standards for implementation. The basic requirement in the domain of critical infrastructure is to identify elements of information security and through interaction with elements of business continuity to confirm the scope and level of functionality in extraordinary circumstances.

## 2. Security mechanism for data transfer

In the realm of data encryption, technological implementation plays a pivotal role in determining the efficacy and security of the measures employed. This intricate landscape involves both commonalities, such as the widespread use of the Advanced Encryption Standard (AES) and global alignment with encryption standards, as well as differences, particularly in the realm of key management. The Advanced Encryption Standard (AES) stands out as a unifying force in data encryption. Its widespread adoption can be attributed to its robust security features, efficiency, and versatility. AES employs symmetric key cryptography, utilizing the same key for both encryption and decryption, making it particularly well-suited for securing data across various applications. Recognized by both governmental and non-governmental entities, AES has become the de facto encryption standard in a multitude of industries, including finance, healthcare, and communication. This commonality ensures a level of interoperability and compatibility across systems, facilitating secure data exchange on a global scale. The importance of global alignment with encryption standards cannot be overstated. Various organizations, including the National Institute of Standards and Technology (NIST) in the United States and the European Union Agency for Cybersecurity (ENISA) in Europe, contribute to the development and promotion of encryption standards. The alignment of these standards on a global scale fosters a common understanding of secure practices, enabling seamless collaboration and communication across borders. This shared commitment to established encryption standards enhances the overall security posture of interconnected systems and promotes a more resilient global information infrastructure. While commonalities exist in encryption algorithms, differences emerge prominently in the realm of key management. Key generation, distribution, and storage are critical aspects that vary across different encryption implementations. Organizations may employ different methods for generating cryptographic keys, ranging from random algorithms to complex mathematical processes. The distribution of keys introduces challenges related to secure transmission and authentication, particularly in large-scale networks. Additionally, the storage of encryption keys demands careful consideration, with options including hardware security modules (HSMs), key management servers, and secure key vaults. The diversity in approaches reflects the nuanced requirements and risk profiles of different industries and applications. The effectiveness of encryption methods is intricately tied to the robustness of key management practices. Weaknesses in key generation, insecure distribution channels, or inadequate storage solutions can compromise the entire encryption system. The impact extends beyond individual organizations to the broader ecosystem, as vulnerabilities in key management can be exploited by malicious actors seeking unauthorized access to encrypted data. Consequently, the effectiveness of encryption methods hinges on the implementation of sound key management practices. Organizations must continually evaluate and enhance their key management strategies to adapt to evolving threats and technological advancements. In conclusion, technological implementation in data encryption is a dynamic interplay of commonalities and differences that shape the security landscape. The widespread use of the Advanced Encryption Standard (AES) and global alignment with encryption standards provide a foundation for secure data practices. However, the varied approaches to key generation, distribution, and storage introduce complexity and demand careful consideration. The effectiveness of encryption methods ultimately relies on the strength of key management practices, emphasizing the importance of ongoing vigilance and adaptation in the face of evolving cyber threats. As technology continues to advance, the harmonization of encryption standards and the refinement of key management strategies will remain essential in safeguarding the integrity and confidentiality of digital information [2].

### 2.1. Security mechanism for data transfer in a virtual environment

Cloud computing technology has gained wide popularity due to its capability to provide enormous resources to individuals and organizations which can be accessed via the internet anytime and anywhere worldwide. Many information and technology (IT) companies have shifted their operations to the cloud, which provides its users with a feature-rich cloud experience, including access to shared resources, which makes resources available when needed at lower costs. These resources may also be swiftly provided and released with minimal administrative effort, and cloud computing provides the ability to

share, manage, and store data, which is actually hosted on remote servers rather than using internal resources or personal devices. Clients can use the cloud services of various programs by adopting cloud computing rather than buying or installing the software on their own computers. Cloud computing provides clients with virtualized resources using various technologies, such as web services, virtualization, applications, and operating systems. The main advantages of cloud computing can be summarized as cost reduction, increased productivity, stability, scalability easy management, and availability. Despite the above advantages of cloud computing, it has given rise to various problems and challenges. Security is one of the greatest hurdles that hinder the acceptance of cloud computing among users. It is a major concern that must be considered, and data security issues arise because client data and software are located on the provider's premises. If suitable security measures are not provided for data operations and transmissions, then they will be at great risk. Thus, cloud service providers (CSPs) must protect data, applications, and cloud infrastructure from internal and external threats. The security of cloud information depends on the implementation of suitable information security measures and countermeasures, making the creation and management of a safe cloud environment a difficult operation. Protecting user data against malicious attacks and unreliable servers is crucial. The user data that have to be secured are: 1) usage data: data gathered from computers; 2) private data, such as bank account and health information; 3) personally identifiable data: data that might be used to define an individual; and 4) unique device identifiers: data that can be uniquely tracked, such as IP addresses. [4] Encryption is one of the safest methods that is used to prevent unwanted access. In cloud computing environments, various types of encryption techniques have been used to protect user data and cloud information, which have contributed to reducing hacking to some extent. The data can be transformed into cipher text to increase their security. However, this process may lead to the loss of numerous features. Encrypted text can be obtained by using two popular techniques. The first technique is based on a safety index, which creates a secure cipher text keyword index by checking the keywords' existence, and the second technique is based on scanning cipher text, which compares each word to ensure that the keywords are in the cipher text. [4]

Information security in a cloud based data center is mostly identical to those of a non-cloud data center. Keeping cloud computing safe from any threats is necessary. Some information security issues that are considered important for a critical infrastructure system are:

1. Malicious insiders: a malicious insider is a person who has authorized access to an organization's network and data and uses these powers in a manner that compromises the organization's information and information systems' confidentiality and integrity. Most organizations are aware of this hazard because it is difficult to detect and has a remarkable effect on the organization.
2. Account or service hijacking: this threat arises because of fraud and software flaws. In this case, an attacker can gain access to sensitive regions on the cloud, in which he can steal permits and sensitive data.
3. Hypervisor vulnerabilities: a hypervisor is the most important piece of software in virtualization. Hypervisors have obvious security vulnerabilities, and remedies are still restricted and often proprietary.
4. Insecure interfaces and application programming interfaces (APIs): If a poor set of interfaces and APIs are used, then organizations may face security threats, such as unknown access, reuse of passwords, the transmission of content or clear text authentication, and inflexible access management or invalid authorizations.
5. Cyber-attacks: hacking and cyber-attacks on networks have increasingly become a great threat in recent years. [4]

In the section Standards in the field of information security and encryption, the authors will present guidelines for solving the mentioned problems by applying list of standards.

Cloud computing provides three key services, namely, software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS). IaaS refers to CSP hardware infrastructure, which includes networks, storage, memory, processors, and a variety of other computing resources. The resources are provided as virtualized systems that can be accessed via the internet. The essential resources are under the control of the CSP. PaaS provides integrated development environments, middleware, operating systems, and platform layer resources through a third-party provider who delivers hardware and software tools to users over the Internet. PaaS does not give customers control over the underlying cloud infrastructure, but only over the applications that are moved to the cloud.

SaaS: It allows consumers to use applications as a service over the internet. Users can simply use the internet to access it rather than buy, install, and maintain software. Customers pay for usage rather than ownership of the software. The cloud computing system is divided into two sections: the front end and the back end. They communicate with each other through a network, usually over the internet. The front end is the side that cloud clients see. The clients do not normally see the back-end section, which includes network connection, cloud servers, and their applications [4]. Next figure shows the categories of cloud services and the architecture of cloud computing.

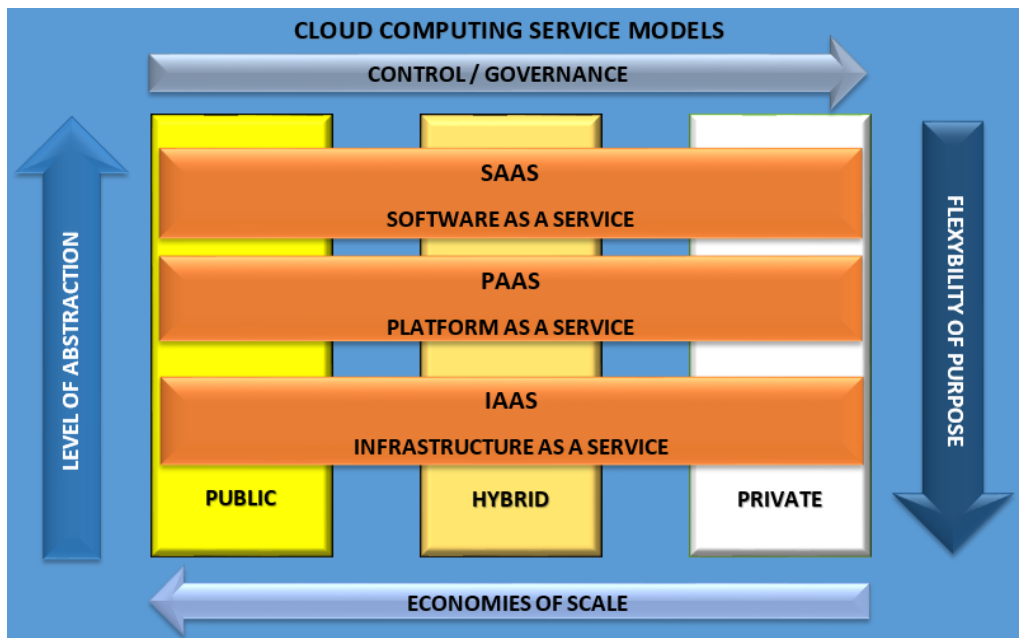


Figure 1. Cloud computing service model [5]

There are different encryption algorithms, with the common ones being symmetric and asymmetric. In symmetric encryption, the same key is applied to encrypt and decrypt. Symmetric algorithms are usually fast and applicable in diverse settings, including encryption of communication links and banking transactions. One of the common symmetric algorithms used today is the Triple Data Encryption Standard (Triple-DES). Triple-DES is an older algorithm, gradually fading and is being replaced with stronger Advanced Encryption Standard (AES). AES uses larger key lengths and encryption block sizes, which makes it stronger, unlike Triple-DES. Further, AES deals with some design issues existing in Triple-DES, which make Triple-DES susceptible to different classes of attacks. In symmetrical encryption, data encryption and decryption happen by splitting it into a form of blocks, which are fed into the cypher system to get an output of cipher text. Due to their ability to handle only small data chunks, symmetric encryption algorithms tend to execute well. One of the significant advantages of symmetrical encryption algorithms is that they are swift, making them ideal for encrypting massive data. Nonetheless, the major drawback of symmetrical key encryption is finding the safest way to exchange the ciphering key with other individuals engaged in the communication channel. This means anyone owning the key can transition the ciphertext to plain text. If a user wants to ensure confidentiality and privacy, keeping the key a secret is vital. Next figure displays the flow path of encrypting and decrypting data, whereby text decryption uses the same symmetric key in decrypting that data back to readable format [3].

Depending upon the key cryptography can be divided into two categories: Symmetric encryption (Private key) and Asymmetric encryption (Public key). Symmetric encryption (private key Encryption) During the encryption and decryption process the same key is used at the sender and receiver site. Before the Transmission of information starts the key distribution has to be made. Example: DES, 3DES, BLOWFISH, AES etc. Asymmetric encryption (Public key encryption) In Asymmetric encryption, two different keys are used for encryption and decryption process. At the same time the two keys are generated. In that one key is transferred to other side before the exchange of information begins. Example: RSA, Elgamal, Elgamal signature Diffie Hellman key exchange, Digital signature [5]

### 3. Critical infrastructure

Critical infrastructure is any system whose interruption of functioning or interruption of delivery of goods or services can have serious consequences on national security, health and lives of people, property, environment, citizens' safety, economic stability, i.e. threatens the functioning of the state. Protection of critical infrastructure is a set of activities and measures aimed at establishing the functioning of critical infrastructure in the event of disruption or destruction, that is, protection in the event of threats and prevention of the consequences of disruption or destruction. The sectors that are most often identified as critical infrastructure are: energy, transport, water and food supply, health, finance, telecommunications and information technologies, environmental protection and the functioning of state bodies. The initial procedural framework for critical infrastructure is the relevant law, most often the Law on Critical Infrastructure and the standard ISO 22301 Security and resilience - Business continuity management systems – Requirements [1].

State bodies are usually charged by law to be guided by the following principles in the area of critical infrastructure protection: 1) the principle of an integrated approach - in the protection of critical infrastructure before, during and after disruptions or interruptions in the functioning of critical infrastructure, competent state authorities and reference entities participate, taking into account the various types of dangers arising from the risk analysis, and taking into account the interdependence of the critical infrastructure sector infrastructures and their interaction; 2) the principle of responsibility - the operators of the critical infrastructure are directly responsible for the functioning of the critical infrastructure, and for the improvement of the protection the critical infrastructure, in addition to the operators, all competent state bodies and reference entities are also responsible; 3) the principle of protection against various types of threats - operators, competent state authorities and reference entities in ensuring the continuous operation of critical infrastructure are obliged to take into account various types of risks; 4) the principle of continuous planning for the protection of critical infrastructure - the protection of critical infrastructure is based on a constant process of risk analysis for the functioning of critical infrastructure and assessment of the adequacy of protection measures; 5) the principle of data and information exchange and data protection - state authorities and reference entities are obliged to timely and continuously exchange the necessary data and information while protecting data related to critical infrastructure, in accordance with the regulations, regulating the protection of secret data.

### 4. Standards in the field of information security and encryption

The basic standard for information security is ISO/IEC 27001 - Information security, cybersecurity and privacy protection - Information security management systems - Requirements. The ISO/IEC 27001 standard provides companies of any size and from all sectors of activity with guidance for establishing, implementing, maintaining and continually improving an information security management system. Conformity with ISO/IEC 27001 means that an organization or business has put in place a system to manage risks related to the security of data owned or handled by the company, and that this system respects all the best practices and principles. In addition to the above-mentioned standard, there are many more standards from the ISO 27K family that additionally define the requirements for the complete implementation of information security, depending on the type of needs or specificity performed by the organization. Annex I of the ISO/IEC 27001 standard in part 8.24 Use of cryptography defines specific requirements in the domain: Rules for the effective use of cryptography, including cryptographic key management, shall be defined and implemented. The implementation of only control 8.24 is not enough, but the author's recommendation is to implement at least the following controls: 8.3 Information access restriction defines specific requirements in the domain: Access to information and other associated assets shall be restricted in accordance with the established topic-specific policy on access control; 8.9 Configuration management defines specific requirements in the domain: Configurations, including security configurations, of hardware, software, services and networks shall be established, documented, implemented, monitored and reviewed; 8.20 Networks security defines specific requirements in the domain: Networks and network devices shall be secured, managed and controlled to protect information in systems and applications; 8.27 Secure system architecture and engineering principles defines specific requirements in the domain: Principles for engineering secure

systems shall be established, documented, maintained and applied to any information system development activities.

The author's recommendation is that during the analysis of the needs for the information security system in critical infrastructure systems, the following standards(for encryption) must be consulted:

- ISO/IEC 15946-1 Information technology - Security techniques - Cryptographic techniques based on elliptic curves Part 1: General. ISO/IEC 15946-1 describes the mathematical background and general techniques necessary for implementing the elliptic curve cryptography mechanisms defined in ISO/IEC 15946-5, ISO/IEC 9796-3, ISO/IEC 11770-3, ISO/IEC 14888-3, ISO/IEC 18033-2 and other ISO/IEC standards. ISO/IEC 15946-1 does not specify the implementation of the techniques it defines. For example, it does not specify the basis representation to be used when the elliptic curve is defined over a finite field of characteristic two. Thus, interoperability of products complying with ISO/IEC 15946-1 will not be guaranteed.
- ISO/IEC 15946-5 Information security - Cryptographic techniques based on elliptic curves Part 5: Elliptic curve generation. The ISO/IEC 15946 series specifies public-key cryptographic techniques based on elliptic curves described in ISO/IEC 15946-1. This document defines elliptic curve generation techniques useful for implementing the elliptic curve based mechanisms defined in ISO/IEC 29192-4, ISO/IEC 9796-3, ISO/IEC 11770-3, ISO/IEC 14888-3, ISO/IEC 18033-2 and ISO/IEC 18033-5. This document is applicable to cryptographic techniques based on elliptic curves defined over finite fields of prime power order (including the special cases of prime order and characteristic two).
- ISO/IEC 18033-1 Information security - Encryption algorithms Part 1: General. This document is general in nature and provides definitions that apply in subsequent parts of the ISO/IEC 18033 series. It introduces the nature of encryption and describes certain general aspects of its use and properties.
- ISO/IEC 18033-2 Information technology - Security techniques - Encryption algorithms Part 2: Asymmetric ciphers. ISO/IEC 18033-2 specifies encryption systems (ciphers) for the purpose of data confidentiality. The primary purpose of encryption (or encipherment) techniques is to protect the confidentiality of stored or transmitted data. An encryption algorithm is applied to data (often called plaintext or cleartext) to yield encrypted data (or ciphertext); this process is known as encryption. The encryption algorithm should be designed so that the ciphertext yields no information about the plaintext except, perhaps, its length. Associated with every encryption algorithm is a corresponding decryption algorithm, which transforms ciphertext back into its original plaintext. An asymmetric, i.e. public-key, encryption scheme allows a sender to use a recipient's public key to transmit an encryption of a message to the receiver, who can use his secret key to decrypt the given ciphertext, thereby obtaining the original message. Such a scheme should be secure in the sense that no information about the message should be leaked to a (resource-bounded) attacker, even if that attacker mounts a so-called 'chosen ciphertext' attack, in which he may obtain decryptions of other ciphertexts. This is the strongest type of attack that has been proposed for a public-key encryption scheme. ISO/IEC 18033-2 specifies the functional interface of such a scheme, and in addition specifies a number of particular schemes that appear to be secure against chosen ciphertext attack.
- ISO/IEC 18033-3 Information technology - Security techniques - Encryption algorithms Part 3: Block ciphers. ISO/IEC 18033 specifies encryption systems (ciphers) for the purpose of data confidentiality. ISO/IEC 18033-3 specifies block ciphers. A block cipher is a symmetric encipherment system with the property that the encryption algorithm operates on a block of plaintext, i.e. a string of bits of a defined length, to yield a block of ciphertext. ISO/IEC 18033-3 specifies following algorithms: 64-bit block ciphers: TDEA, MISTY1, CAST-128, HIGHT; 128-bit block ciphers: AES, Camellia, SEED.
- ISO/IEC 18033-4 Information technology - Security techniques - Encryption algorithms Part 4: Stream ciphers. ISO/IEC 18033-4 specifies output functions to combine a keystream with plaintext, keystream generators for producing keystream, and object identifiers assigned to dedicated keystream generators in accordance with ISO/IEC 9834.
- ISO/IEC 18033-5 Information technology - Security techniques - Encryption algorithms Part 5: Identity-based ciphers. ISO/IEC 18033-5 specifies identity-based encryption mechanisms. For each

mechanism the functional interface, the precise operation of the mechanism, and the ciphertext format are specified.

- ISO/IEC 18033-6 IT Security techniques - Encryption algorithms Part 6: Homomorphic encryption. This document specifies the following mechanisms for homomorphic encryption: Exponential ElGamal encryption; Paillier encryption. For each mechanism, this document specifies the process for: generating parameters and the keys of the involved entities; encrypting data; decrypting encrypted data; and homomorphically operating on encrypted data. Annex A defines the object identifiers assigned to the mechanisms specified in this document.
- ISO/IEC 18033-7 Information security - Encryption algorithms Part 7: Tweakable block ciphers. This document specifies tweakable block ciphers. A tweakable block cipher is a family of  $n$ -bit permutations parametrized by a secret key value and a public tweak value. Such primitives are generic tools that can be used as building blocks to construct cryptographic schemes such as encryption, Message Authentication Codes, authenticated encryption, etc. A total of five different tweakable block ciphers are defined and they are categorized.
- ISO/IEC 19772 Information security - Authenticated encryption. This document specifies five methods for authenticated encryption, i.e. defined ways of processing a data string with the following security objectives: data confidentiality, i.e. protection against unauthorized disclosure of data; data integrity, i.e. protection that enables the recipient of data to verify that it has not been modified; data origin authentication, i.e. protection that enables the recipient of data to verify the identity of the data originator. All five methods specified in this document are based on a block cipher algorithm, and require the originator and the recipient of the protected data to share a secret key for this block cipher.
- ISO/IEC 29150 Information technology - Security techniques - Signcryption. ISO/IEC 29150 specifies four mechanisms for signcryption that employ public key cryptographic techniques requiring both the originator and the recipient of protected data to have their own public and private key pairs. The methods specified in ISO/IEC 29150 have been designed to maximize the level of security and provide efficient processing of data. All the mechanisms defined have mathematical "proofs of security", i.e. rigorous arguments supporting their security claims.
- ISO/IEC 29192-1 Information technology - Security techniques - Lightweight cryptography Part 1: General. ISO/IEC 29192-1 provides terms and definitions that apply in subsequent parts of ISO/IEC 29192. ISO/IEC 29192-1 sets the security requirements, classification requirements and implementation requirements for mechanisms that are proposed for inclusion in subsequent parts of ISO/IEC 29192.
- ISO/IEC 29192-2 Information security - Lightweight cryptography Part 2: Block ciphers. This document specifies three block ciphers suitable for applications requiring lightweight cryptographic implementations: PRESENT: a lightweight block cipher with a block size of 64 bits and a key size of 80 or 128 bits; CLEFIA: a lightweight block cipher with a block size of 128 bits and a key size of 128, 192 or 256 bits; LEA: a lightweight block cipher with a block size of 128 bits and a key size of 128, 192 or 256 bits.
- ISO/IEC 29192-3 Information technology - Security techniques - Lightweight cryptography Part 3: Stream ciphers. ISO/IEC 29192-3 specifies two dedicated keystream generators for lightweight stream ciphers: Enocoro: a lightweight keystream generator with a key size of 80 or 128 bits; Trivium: a lightweight keystream generator with a key size of 80 bits.
- ISO/IEC 29192-4 Information technology - Security techniques - Lightweight cryptography Part 4: Mechanisms using asymmetric techniques. ISO/IEC 29192-4 specifies three lightweight mechanisms using asymmetric techniques: a) a unilateral authentication mechanism based on discrete logarithms on elliptic curves; b) an authenticated lightweight key exchange (ALIKE) mechanism for unilateral authentication and establishment of a session key; c) an identity-based signature mechanism.
- ISO/IEC 29192-5 Information technology - Security techniques - Lightweight cryptography Part 5: Hash-functions. ISO/IEC 29192-5 specifies three hash-functions suitable for applications requiring lightweight cryptographic implementations.- PHOTON: a lightweight hash-function with permutation sizes of 100, 144, 196, 256 and 288 bits computing hash-codes of length 80, 128, 160, 224, and 256 bits, respectively. - SPONGENT: a lightweight hash-function with permutation sizes



of 88, 136, 176, 240 and 272 bits computing hash-codes of length 88, 128, 160, 224, and 256 bits, respectively. - Lesamnta-LW: a lightweight hash-function with permutation size 384 bits computing a hash-code of length 256 bits.

- ISO/IEC 29192-6 Information technology - Lightweight cryptography Part 6: Message authentication codes (MACs). This document specifies MAC algorithms suitable for applications requiring lightweight cryptographic mechanisms. These mechanisms can be used as data integrity mechanisms to verify that data has not been altered in an unauthorized manner. They can also be used as message authentication mechanisms to provide assurance that a message has been originated by an entity in possession of the secret key. The following MAC algorithms are specified in this document: a) LightMAC; b) Tsudik's keymode; c) Chaskey-12.
- ISO/IEC 29192-7 Information security - Lightweight cryptography Part 7: Broadcast authentication protocols. This document specifies broadcast authentication protocols, which are protocols that provide data integrity and entity authentication in a broadcast setting, i.e. a setting with one sender transmitting messages to many receivers. To provide entity authentication, there needs to be a pre-existing infrastructure which links the sender to a cryptographic secret. The establishment of such an infrastructure is beyond the scope of this document.
- ISO/IEC 29192-8 Information security - Lightweight cryptography Part 8: Authenticated encryption. This document specifies one method for authenticated encryption suitable for applications requiring lightweight cryptographic mechanisms. This method processes a data string with the following security objectives: a) data confidentiality, i.e. protection against unauthorized disclosure of data, b) data integrity, i.e. protection that enables the recipient of data to verify that it has not been modified. Optionally, this method can provide data origin authentication, i.e. protection that enables the recipient of data to verify the identity of the data originator.

## 5. Conclusions

Information security is an indispensable point of analysis of all information systems, especially information systems recognized by law as critical infrastructure systems. Encryption of particularly sensitive parts of information systems is important because the task is to ensure that certain information is not accessible to everyone in the internal or external system. The paper gave recommendations for the implementation of the encryption system in information security based on the author's experience in numerous implementations of the ISO/IEC 27001 standard in critical infrastructure systems in the countries of Southeast Europe. Special attention must be paid to issues of treatment and risk analysis for each system, method or applied solution. The specific recommendations are technological guidelines for information technology engineers who are tasked with establishing an adequate encryption system in all critical infrastructure systems.

### References:

- [1] Akoh Atadoga, Oluwatoyin Ajoke Farayola, Benjamin Samson Ayinla, Olukunle Oladipupo Amoo, Temitayo Oluwaseun Abrahams, & Femi Osasona (2024). A comparative review of data encryption methods in the USA and Europe. Computer Science & IT Research Journal, P-ISSN: 2709-0043, E-ISSN: 2709-0051, Volume 5, Issue 2, P.447-460, February 2024, DOI: 10.51594/csitrj.v5i2.815,
- [2] Yahia Alemami, Ali M. Al-Ghonmein, Khaldun G. Al-Moghrabi, Mohamad Afendee Mohamed, Cloud data security and various cryptographic algorithms, International Journal of Electrical and Computer Engineering (IJECE) Vol. 13, No. 2, April 2023, (pp. 1867~1879) ISSN: 2088-8708, DOI: 10.11591/ijece.v13i2.pp1867-1879
- [3] Sandeep Reddy Gudimetla, Data encryption in cloud storage, International Research Journal of Modernization in Engineering Technology and Science, Volume:06/Issue:03/March-2024
- [4] Dragović, R., Dobrilovic, D., & Dragović, D. (2022). Recommendations for the Creation of Usable Critical Infrastructure for the Delivery of Priority Services of State Bodies. In IFIP International Conference on Human Choice and Computers (pp. 161-171). Cham: Springer Nature Switzerland. DOI:10.1007/978-3-031-47990-8\_15,

- [5] Tangirala, Suryakanthi. "Efficient big data analytics and management through the usage of cloud architecture." *Journal of Advances in Information Technology* Vol 7.4 (2016).
- [6] Dragović, R., Stanisavljev, S., Dobrilovic, D., Dragović, D., Milosev, V., *Software information security management for government authorities*, IIZS 2023, ISBN 978-86-7672-368-3

# Interactive faculty map and timetable web application

Ana Marija Ognjenovic<sup>1</sup>, Bosko Nikolic<sup>1</sup>, Visnja Ognjenovic<sup>2</sup>

<sup>1</sup> School of Electrical Engineering, University of Belgrade, Belgrade, Serbia

<sup>2</sup> Technical Faculty "Mihajlo Pupin" Zrenjanin, University of Novi Sad, Serbia

anamognjenovic@gmail.com; nbosko@etf.bg.ac.rs; visnjao@tfzr.uns.ac.rs

## Abstract:

This paper presents the development of an interactive faculty map and timetable web application, named eRaspored, aimed at enhancing the schedule handling for students and professors at the School of Electrical Engineering, University of Belgrade. Utilizing the MEAN stack (MongoDB, Express, Angular, Node.js), the application offers a user-friendly interface for managing class schedules, including features such as an interactive map for classroom visualization and schedule search functionality based on various criteria. The design addresses common scheduling challenges, such as resource allocation and availability constraints. This paper also evaluates existing scheduling solutions, discusses the technologies used, and suggests potential improvements, with an emphasis on the automation of schedule generation.

## Keywords:

class schedule, MEAN stack, web app

## 1. Introduction

A class schedule is a simple yet crucial tool that helps students and professors align their responsibilities and structure their day. Despite its simplicity, it plays a practical and important role in everyone's daily routine at school or university. While a traditional schedule can be effective, utilizing modern technologies can elevate it to a whole new level.

In essence, class scheduling revolves around managing several key parameters, such as professors, subjects, student groups, classrooms, days and classes. These elements must be carefully coordinated to ensure that the schedule aligns with both the academic needs and available resources.

However, scheduling becomes more complex when dealing with special cases. For instance, scenarios where multiple student groups need to attend the same class together must be carefully coordinated within the schedule. Some other challenges that could possibly arise while creating the schedule include:

- Professor A teaches a class in the laboratory X, therefore it is not available for professor B
- A professor may be available only on certain days of the week
- ... and many other constraints and requirements

These problems could be solved programmatically in a variety of efficient ways. Designing the schedule requires balancing key factors like teacher availability, classroom allocation, and group coordination to create an organized and effective schedule. Once in use, the schedule ensures that daily activities run smoothly, preventing conflicts and maximizing productivity. A well-crafted schedule is not just a plan, it is an essential tool that enables students and professors to navigate their daily responsibilities with ease.

This paper is divided into six chapters. The first chapter introduces the subject, outlining the general requirements of the application and the key concepts involved in timetable management. Within Chapter 2, existing scheduling software are reviewed. In Chapter 3, a brief overview of the technologies used in the development of the system is shown. Chapter 4 describes eRaspored application itself, detailing its features and functionalities. Chapter 5 explores potential improvements and optimizations, such as the possibility of automating schedule generation. Finally, the last chapter concludes the paper by summarizing the key findings and discussing the overall significance of the application.

## 2. Previous work

There are many existing software for creating, managing and viewing class schedules. Some have the ability to automatically generate schedules, taking into account various rules, whereas others are much more simple and rely on the manual approach. Described below are a few interesting examples of timetable-generating software, as well as ones that served as inspiration for this web application.

### 2.1. DEDUC

DEDUC is a software program designed for efficient class schedule generation, developed in 2001. at the Technical Faculty “Mihailo Pupin” in Zrenjanin under the guidance of professor Petar Hotomski [1]. It is based on theoretical foundation of the Automated Theorem Proving, utilizing the Resolution principle and heuristic methods of the Artificial Intelligence [2]. The main focus of this program is efficient generation of class schedules, while adhering to user-defined requirements. This program offers a wide range of capabilities, allowing detailed customization of parameters and requirements for automatic schedule generation.

The system offers several key features:

- Automatic and interactive schedule generation based on initial data and specified conditions
- A user-friendly interface for managing and configuring initial data and requirements
- Flexibility to cater to various needs, such as splitting classes into groups, linking them, setting restrictions on the use of labs and classrooms, and accommodating specific teacher preferences
- Support for generating schedules that consider different school shifts
- Convenient browsing and printing schedules by class or teacher
- Capability to generate and archive multiple schedule versions using the same data and conditions
- Maintenance of the archived versions, data and conditions

If contradictions are present within the requirements, timetable generation is halted. The system allows for the adjustment of requirements to ensure successful creation of the class schedule. Real-world experiences show that the DEDUC system produces timetables with high level of quality, surpassing those made by hand. The time required for generating a timetable ranges from a few minutes to several hours, depending on the volume of data, complexity of the conditions and computer performance.

### 2.2. ascTimetables

ascTimetables [3] is a scheduling software designed for creation and management of class schedules in educational institutions. Its user-friendly interface allows users to work with various parameters, such as subjects, teachers, classrooms and student groups. The program is designed to simplify the scheduling process and to ensure that all requirements are met. By automating the scheduling workflow, ascTimetables allows educators to save valuable time and effort, taking on mundane administrative tasks.

In this program, users can easily add subjects, which are scheduled into the class timetable. For each class, specific time slots can be assigned and groups can be formed by dividing a class. The formation of groups allows merging and separating based on the requirements. User has the ability to set restrictions and add rules for teachers. These rules may include limits on the maximum number of periods they can teach, the duration of breaks between sessions, and the total periods permitted each day. In addition, users can define parameters such as the maximum number of consecutive periods and the allowable number of free periods each week, ensuring a balanced workload for teachers.

The process of generating the timetable begins with an empty schedule and an organized panel displaying unallocated periods. Using provided data and set restrictions, the program automatically creates a complete timetable. The final schedule is showcased in the timetable panel, while any unallocated periods remain visible in their dedicated section. A successful generation of the timetable is indicated by the absence of unallocated period cards, as shown in Figure 1.

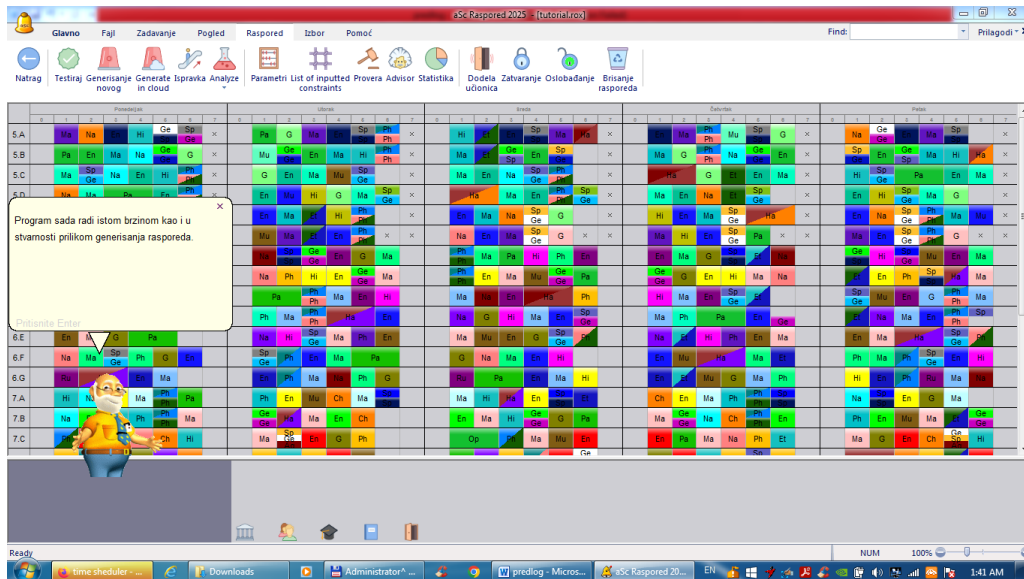


Figure 1: Successfully generated schedule

This program allows users to display schedules for specific class, student, teacher, classroom and subject. For fine tuning and last minute changes, it has the ability for manual adjustment of individual periods. If any scheduling conflicts are detected, the software alerts the user. Finally, the generated timetable can be tested, and the program informs the user of the results of the verification process, indicating any violated rules or unmet conditions.

## 2.3. eStudent

eStudent[4] is an online portal used by educational institutions to provide a range of services and tools for addressing student inquiries, from which the application derives its name, eRasposed. This site offers various functionalities, but the relevant feature here is the schedule overview. Users need to login to gain access. In the schedule section, students are shown their default schedule, but there are also options to search for schedules by student, teacher, subject, department, and classroom. As users begin to enter search parameters, a dropdown list appears with suggestions that update in real time based on the entered data. The interface is simple yet effective, allowing for simple search of schedules based on the listed criteria.

## 3. Used technologies

This application was developed using several technologies that together form the MEAN stack (MongoDB, Express, Angular, Node.js) [5]. This combination provides a solid foundation for building responsive, scalable and dynamic applications, while ensuring that the code remains modular, clean and well-organized.

The frontend of the application is built using Angular platform. Its component-based architecture, two-way data binding and vast ecosystem make Angular a preferred choice for developing modern and scalable web applications. Programming languages used for implementation of the interactive elements of this app are TypeScript, CSS and HTML.

On the backend side of this application, Node.js and Express are used for server-side functionalities, working together with MongoDB database for data storage. Node.js is a platform designed to process user requests, while Express is a framework built on Node.js that simplifies the creation of APIs for developers. MongoDB is a non-relational database that allows flexible data storage and retrieval, making it well-suited for apps that require scalability and high performance.

#### 4. Description of the developed app

The eRaspored web application is an online platform that allows students and professors to view, search and manage class schedules. The application is designed to offer quick and easy access to different schedules, while ensuring smooth and intuitive user experience. A standout feature of this application is its interactive faculty map, which provides users with a visual representation of the space and enables quick identification of the available classrooms, or classrooms with specific purposes. This intuitive visual tool enhances the user experience by making it easier to navigate through the faculty.



Figure 2: Main page

The map feature offers users a visual representation of classroom layouts, showing a detailed map of the ground floor and third floor of the Technical Faculty building in Belgrade. The ground floor is shown by default, as illustrated in Figure 2. Users can easily switch between floors with a single click using the floor selection option located in the side menu. While the map is not essential for the core functionality of the scheduling application, it provides an intuitive and visually appealing navigation experience through the faculty building. By clicking on a specific room on the map, the user can view the detailed schedule for that particular classroom, as shown in Figure 3. This feature provides quick and easy access to the schedule, allowing users to instantly see when a classroom is free or occupied.

Vreme	Ponedeljak	Utorak	Sreda	Četvrtak	Petak
08:00					
09:00					
10:00					
11:00			Programiranje Internet Aplikacija - Biljana Ivanović [70]		
12:00				Programiranje Internet Aplikacija - Dunja Marić [70]	
13:00		Sistemske Softver - Marija Marković [70]			
14:00					
15:00					
16:00				Sistemske Softver - Marija Marković [70]	
17:00					
18:00					
19:00					
20:00					

Figure 3: Classroom schedule

An important functionality of this application is the ability to search schedules based on specific criteria. The available search options include student, professor, subject and department. By clicking the search button in the header, a modal window is shown where user selects a criterion and enters necessary information. The layout of this modal window can be seen in Figure 4. For a student's schedule search, index has to be inputted, while the professors are chosen from a sorted drop-down list. Department timetable search requires selectiong both the department and the academic year. Timetable search is an essential as it simplifies the search process based on various criteria, enabling students and professors to easily track their schedules, while simplifying schedule management for the admin.

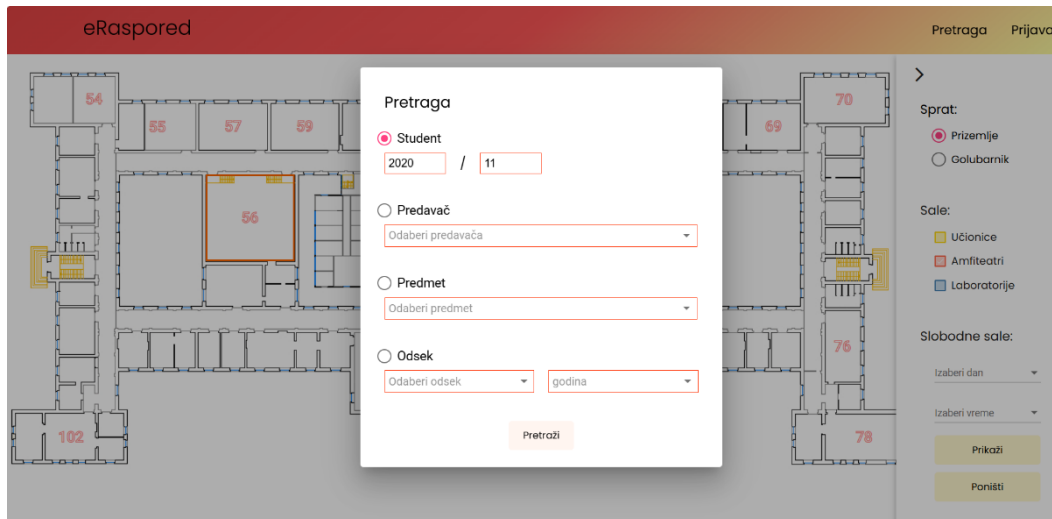


Figure 4: Schedule search

The user can search for available rooms for a specific time slot by selecting the start time and the day of the week from the dropdown menus. As shown in Figure 5, once the information is entered, the map highlights the classrooms that are free during that period, marking them in green, while the occupied ones remain unmarked. With this feature, available spaces are easily found.



Figure 5: Searching for available classrooms

The option to highlight different types of classrooms on the map allows users to quickly locate a specific type of classroom. They can mark standard classrooms, amphitheatres and laboratories in yellow, red and blue, respectively, by selecting the appropriate checkboxes from the side menu. This feature is particularly useful for administrators when scheduling a class that requires specific conditions, such as large or small number of students, or classes where students need to use computers. Highlighting different types of rooms can be seen in Figure 6.



Figure 6: Selection of classroom types

The login feature, accessible via the login button in the header, is specifically designed for administrators. They must enter their username and password to gain access. Once logged in, the administrator can modify the schedule by adding and deleting periods within the schedules. By clicking on an empty time slot within the schedule, admin can add a period for a specific subject. From a dropdown menu, a specific subject and professor are selected, as shown in Figure 7. If the selected professor is not occupied during that time, the period addition is successful. By clicking on an occupied time slot in the schedule, the administrator can see period details, shown in Figure 8. To remove a period, administrator has to click on the occupied time slot in the schedule and then click on the delete button. The deletion action is always available.

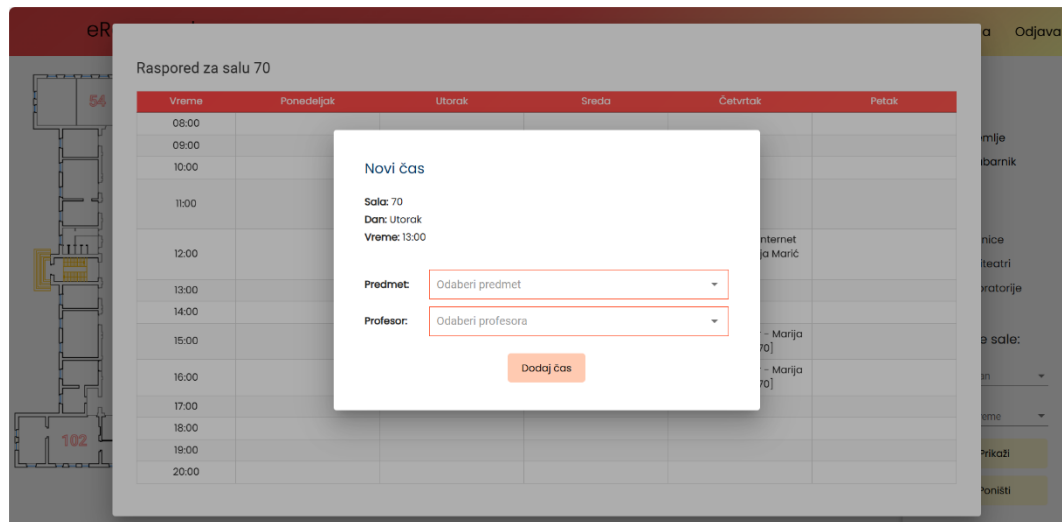


Figure 7: Adding a new period



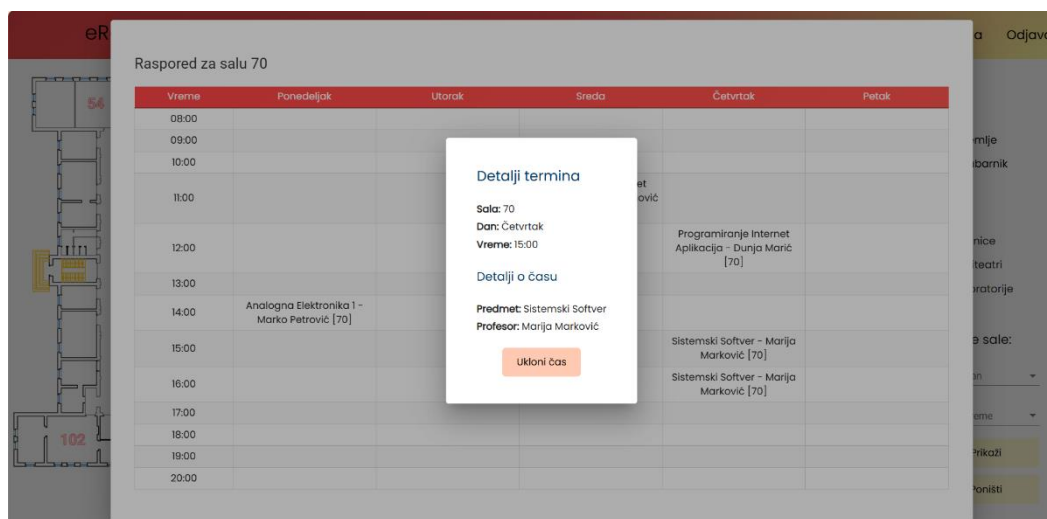


Figure 8: Period details

## 5. Discussion

Automated schedule generation provides significant advantages in efficiency and accuracy. It excels at handling complex requirements, such as professor availability and classroom restrictions, which can result in high-quality schedules produced in a fraction of the time required for manual processes. This approach reduces the potential for conflicts and optimizes resource allocation. However, there are situations where manual scheduling may be more effective. For instance, when dealing with special events, last-minute changes, or unique requirements that require careful consideration of the context, human oversight can be crucial. In these cases, the flexibility afforded by manual scheduling allows for adjustments that automated systems may not be able to address adequately.

Creating timetables is the most challenging aspect of managing timetables, offering a potential for optimization. Future development should improve eRaspored by integrating automated schedule generation techniques, similar to those used in DEDUC and ascTimetables, to improve efficiency and assist scheduling creation, with algorithms that analyze historical data to optimize scheduling patterns. Additionally, adding a feature to distinguish between professors and assistants, lectures, and exercises could improve the user experience.

## 6. Conclusion

In this paper, the development of the eRaspored web application was presented, designed to enhance the schedule management at the School of Electrical Engineering, University of Belgrade. A class schedule serves as an essential tool that helps students and professors align their responsibilities and structure their day. Although traditional timetables are useful, modern technologies can greatly enhance the overall experience. Made using MEAN stack technologies, the eRaspored application provides an intuitive platform for users to view, search and manage schedules through an interactive map. The application allows users to search schedules by a given student, professor, department and classroom, as well as to find available classrooms for a particular time slot. A visually dynamic faculty map allows users to quickly access classroom schedules by clicking on each classroom, and includes color-coded markers for displaying different types of classrooms on the faculty.

At present, the application offers an interface for viewing and searching for schedules for common users: students and professors. Furthermore, it offers functionalities that allow the administrator to manually change the schedule, by adding and removing periods. Manual schedule creation is precise but also an effortful and time-consuming task. Therefore, potential future enhancements, such as the automation of schedule generation, could further improve efficiency and accuracy in schedule management. Automated schedule generation, as found in tools like DEDUC and ascTimetables, has the potential to reduce manual labor, optimise available space, and create conflict-free schedules.

**References:**

- [1] SYSTEM FOR GENERATING CLASS SCHEDULES, 2002. URL: [https://deduce.tripod.com/info\\_bas.htm](https://deduce.tripod.com/info_bas.htm)
- [2] Berković, Ivana F. "ORDERED LINEAR RESOLUTION AS THE BASE OF THE SYSTEM FOR AUTOMATIC THEOREM PROVING." *Filomat*, vol. 9, no. 3, 1995, pp. 591–97. *JSTOR*, <http://www.jstor.org/stable/43999238>. Accessed 28 Oct. 2024.
- [3] ascTimetables, 2024. URL: [https://www.asctimetables.com/?gad\\_source=1&gclid=Cj0KCQjwgrO4BhC2ARIsAKQ7zUmpkQ9Q6jQ10O6vFgCmexS3PkW5s5P5hO1xaXrDZ4Ty8Fi1I4ALAlaAneiEALw\\_wcB#!/home/info](https://www.asctimetables.com/?gad_source=1&gclid=Cj0KCQjwgrO4BhC2ARIsAKQ7zUmpkQ9Q6jQ10O6vFgCmexS3PkW5s5P5hO1xaXrDZ4Ty8Fi1I4ALAlaAneiEALw_wcB#!/home/info)
- [4] eStudent, 2024. URL: <https://student.etf.bg.ac.rs/pocetna.jsf>
- [5] Ana Marija Ognjenović, Veb aplikacija za rad sa interaktivnom mapom fakulteta i rasporedom časova, Diploma Thesis, University of Belgrade, Belgrade, 2024.

# Development of a Responsive Web Application to Support the Creation of Class Schedules

Ivana Samardzic<sup>1</sup> and Dragoljub Krneta<sup>1</sup>

<sup>1</sup> University of East Sarajevo, East Sarajevo, Republic of Srpska, Bosnia and Herzegovina

ivana.sipovac@ffuis.edu.ba; d.krneta@ffuis.edu.ba

## Abstract:

The motivation for writing this paper is to find an adequate software solution that would increase the productivity and efficiency of employees responsible for creating class schedules. One of the main goals is to utilize modern technologies in the development of a responsive web application, which, in addition to a well-structured and normalized database, will have a high-quality and user-friendly interface suitable for use on various devices. The web application needs to meet the requirements of responsive design, meaning it should be usable on screens of different sizes.

## Keywords:

Web, Application, Responsiveness, Class, Schedules

## 1. Introduction

The class schedule represents the timetable of subjects that a student or group of students can attend during a working day or a week. Each schedule includes the subject name, date, exact time, and the name of the classroom where the class will take place. Additionally, the schedule contains the name of the teacher, professor, or assistant who will conduct the class. The schedule serves as a tool for students and professors to manage their time as efficiently as possible. This paper will present a solution for developing a web application for creating schedules using modern environments and tools.

Web programming is a field of programming that focuses on the development of software solutions using web technologies. Such solutions are used in a web-based environment, implying the use of appropriate network infrastructure and the Internet. In a broader context, web programming can also be viewed as one segment within web design as a multidisciplinary field that, in addition to programming, encompasses graphic design, user interface design, information structure design, content organization of web documents, and working with multimedia content such as images, audio, and video [1].

A web application is application software that resides on a web server. Web applications are run from a web browser and are not dependent on the operating system or the type of device through which the end user accesses the application [2]. The implementation and maintenance of web applications occur on the web server, rather than on user computers, which can be a significant advantage in organizations and systems with a large number of users. Any code changes to the web application on the production server, or updates to the database, immediately become available to all users.

The development of web applications is typically based on a three-tier architecture. The three-tier architecture is a well-established software application architecture that organizes applications into three logical and physical computing layers: the presentation layer, the application layer, and the data layer. The presentation layer is the user interface and communication layer, where the end user interacts with the application. The presentation layer is usually developed using HTML, CSS, and JavaScript. The application layer is also known as the logic layer or middle layer. In this layer, the information gathered from the presentation layer is processed using business logic, a specific set of business rules. The application layer is typically developed using PHP, Java, C#, Python, and others. The data layer is where the information processed by the application is stored and managed. This can be a relational database management system such as PostgreSQL, MySQL, MariaDB, Microsoft SQL Server, and others [3].

To develop a web application that meets the specified scheduling functionalities, the following technologies were used:

- HTML, CSS, and Vaadin for creating the user interface and responsive design;
- Java programming language for developing the application layer;

- MySQL database management system and SQL (Structured Query Language) for data management.

SQL is a language designed for querying and managing data in relational database management systems, for creating and modifying relational database schemas, and for controlling access to specific database objects [4].

HTML (HyperText Markup Language) is used for creating web pages and is characterized by tags that separate instructions from content. CSS (Cascading Style Sheets) is a language used for styling the appearance of one or more elements on a web page, or the web page as a whole.

Vaadin is an open-source platform for building modern web applications for Java backends. All Vaadin components are designed primarily for mobile devices and adapt to different screen sizes. Custom CSS files and class names can be added to any component. Vaadin-based applications can also be listed in the Windows Store. There are two ways to create a Vaadin application: the Flow and Hilla frameworks [5].

During the development of the application, the following software and programming environments were used:

- <https://app.creately.com/> - for creating use case diagrams,
- <https://online.visual-paradigm.com/> - for creating sequence diagrams,
- Power Designer (trial version) – for designing the physical database model,
- MySQL – database management system,
- Eclipse IDE for Java Developers – 2023 – for developing the application layer of the web application,
- Vaadin – open-source platform for creating user interfaces.

CASE tools are used today for database modelling. A CASE product is any software product designed to support or automate at least one task within the lifecycle of another software product, or to provide complete support for the design and implementation of another software product.

## 2. Related Works

Historically, the University class scheduling problem has been difficult to solve efficiently, due to the long time it takes to find satisfactory solutions [6]. Timetabling problem is known as an NP-hard problem that centres around finding an optimized allocation of subjects onto a finite available number of slots and spaces. It is perhaps the most challenging issues looked by colleges around the globe. Every academic institution faces a problem when preparing courses and exam plans. There are many restrictions raised while preparing a timetable [7]. Making a class schedule for next semester that suits a student's interests and meets overall graduation requirements within a time frame is not always easy. The changing requirements, transferable units of different schools, the availability of classes and other factors can often cause mistakes in the scheduling process. An automated tool can help students to identify scheduling mistakes and compare available options [8]. In the age of mobile technology, the demand for seamless access to educational resources has become essential [9].

In article [10] describes the process of creating a class schedule in educational institutions with a classical education system. The requirements and restrictions for the class schedule are given. A model for interactive scheduling of classes in an educational institution is proposed. Interactive scheduling, which is based on the “traffic light” principle, frees the user from the control of all hard restrictions and warns about violations of soft restrictions, which ensures that there are no conflict situations in the schedule. The work [9] presents a project the integration of a Learning Management System (LMS) and a class schedule platform of a higher education institution, developing for that a mobile application aimed at integrating with the LMS used at the institution. The application seeks to enhance the learning experience by providing students, instructors, and administrators with anytime, anywhere access to the LMS core functionalities.

The paper [8] introduces a web-based expert system, called class schedule planner (CSP). CSP encapsulates class-scheduling knowledge and gives intelligent scheduling advises to students. It uses JESS (Java expert system shell) as a tool to process the facts and rules to generate feasible schedules for the students.

### 3. Development of the Solution

The development of a responsive web application to support class schedule generation was carried out through the following phases: user requirements analysis and presentation, database model creation, application layer development, user interface design, application testing, and the resolution of identified bugs or undesired application behaviours.

#### 3.1. User Requirements

The first step in the software development process is documenting the requirements that the software needs. The set of requirements arises as a result of an analysis that should be conducted to understand the fundamental problems and needs of the client. This analysis involves intensive collaboration with clients, particularly with the software users. The success of this analysis greatly influences the outcome of the entire software development project. A requirement represents an expression of the desired behavior of the software.

When defining requirements, the characteristics of objects and entities present in the system are considered, the states in which these objects and entities may exist, as well as the functions that enable changes to the states or characteristics of the objects. The goal of requirements analysis is to precisely establish what behavior the client expects from the software. At this stage, there is no consideration of how this behavior will be implemented in the system. This means that the choice of technology for system realization, the selection of a database, or the architecture of the system to be used are not discussed at this point. During the analysis, requirements are typically expressed through descriptions from the client's real-world environment, without using technical jargon that will be employed in the system [11].

When defining the requirements for the application development, the following functionalities were specified:

- **Data Entry:** The ability to input subjects, teachers, students, and available classrooms and time slots.
- **Schedule Generation:** Automatic generation of schedules based on defined rules (e.g., ensuring classes do not overlap).
- **Schedule Display:** An interactive display of the schedule (e.g., calendar or table format).
- **Modifications and Customization:** The option for manual changes to the schedule.
- **Download and Sharing:** Options to download the schedule in various formats (PDF, Excel) and share it via email.
- **User Interface:** Responsive design to support various devices (desktop, tablet, mobile), along with intuitive navigation and ease of use.

Users should access the solution from four different perspectives, meaning there are four distinct roles with different permissions and privileges:

- **Professor/Assistant:** A user who can view the schedule.
- **Student:** Has identical privileges as the professor/assistant.
- **Secretary:** A user who can create lecture schedules for a specific department (adding, deleting, and modifying lectures in the schedule) and view lecture schedules.
- **Administrator:** A user with the same authority as the secretary, who can make changes to the database, including adding, modifying, and deleting information about department names, subjects, and personal details of professors and assistants, among others.

We will use use case diagrams and sequence diagrams to graphically represent the user requirements. UML (Unified Modeling Language) is a standardized visual language used for modeling software systems. With UML, UML diagrams can be created—graphic representations used for modeling software support. They provide a standardized and visual way to represent the architecture, design, and behavior of complex software systems, facilitating communication, collaboration, and understanding among development engineers.

Use case diagrams are one of the most commonly used types of diagrams in software engineering. Their primary purpose is to visually model the functionality of the system and depict the interactions between various actors and the system itself. Due to their simplicity and high level of abstraction, use case diagrams are easy to understand even for stakeholders who are not professional software engineers (such as clients, team members, and managers). They help in analyzing and understanding the relationships between different functionalities and actors, as well as in identifying potential areas for improvement or missing functionalities [12].

Figure 1 presents the use case diagram for the previously described user requirements.



Figure 1. Use Case diagram

Use case diagrams provide a basic representation of the interaction between external actors and the system, as well as the specific functionalities that the system offers. However, to better understand the details of these interactions, such as the sequence of events or conditional execution, additional information is needed. Use case diagrams alone are not expressive enough. To achieve a more detailed visualization and analysis of the dynamic behavior of the system, sequential UML diagrams are used. UML sequence diagrams are behavior diagrams that illustrate the order of interactions, or message exchanges, between objects in the system over time. They are semantically richer than use case diagrams, include a broader set of elements, and have a somewhat more complex syntax [12].

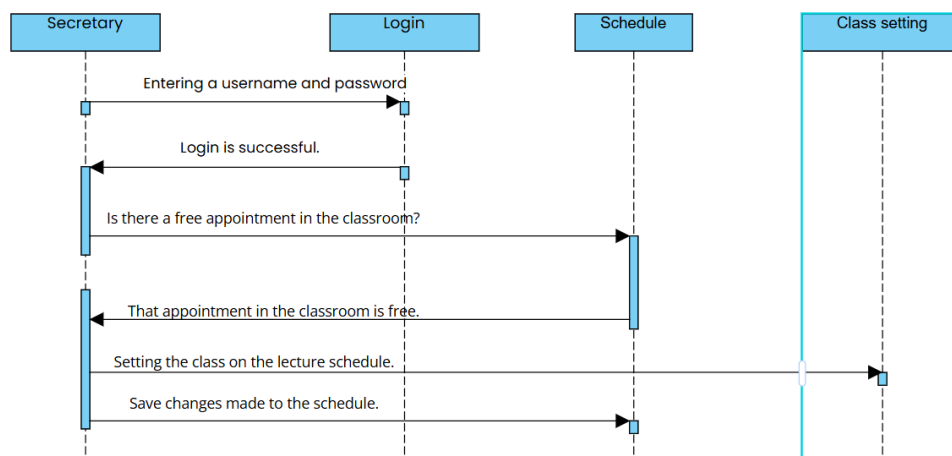


Figure 2. Sequence diagram

It is important to emphasize the time required to perform a given activity. The following figure is a sequence diagram representing the activity of scheduling classes by the secretary of a specific department. To execute this action, the secretary must first successfully log in by entering the correct

username and password. Additionally, time is needed to find an available time slot in a specific classroom, and afterwards, it is also necessary to save those changes.

### 3.2. Database Design

The data model is a set of rules that defines how the logical structure of a database can look. A data model is a specific theory used to design a database. Modelling is the process of using a model instead of the real system. The goal of modelling is to create a clear and realistic representation of the system, or to visualize the system. A view of the database in which certain details are intentionally ignored is called abstraction. An entity is a concept that can be clearly identified and represents everything about which the database management system needs to store certain data.

According to the level of abstraction, models can be divided into conceptual, logical, and physical models. The model with the highest level of abstraction is the conceptual model, while the model with the lowest level of abstraction is the physical model. Modelling is carried out in parallel with requirements analysis. As information is gathered, objects are identified, and names are assigned to them using terms familiar to end users [13].

Regardless of their role, all application users have personal data: first name, last name, and login information. However, not all users have the same rights. User permissions, depending on the module, may include viewing, adding, deleting, and modifying content. For locations, departments, semesters, and classrooms, only their names are essential. Teaching staff have personal data, including first name, last name, contact phone number, unique citizen identification number, address, information about the department they belong to, and whether they serve as an assistant or a professor at the faculty. For courses, the most important data includes the course name, information about the semester in which the course can be taken, and the number of ECTS credits that the course carries. The schedule contains data about courses, dates, times, duration of lectures, the classroom where classes are held, and information about the teaching staff conducting the classes. Figure 3 shows the physical model of the database.

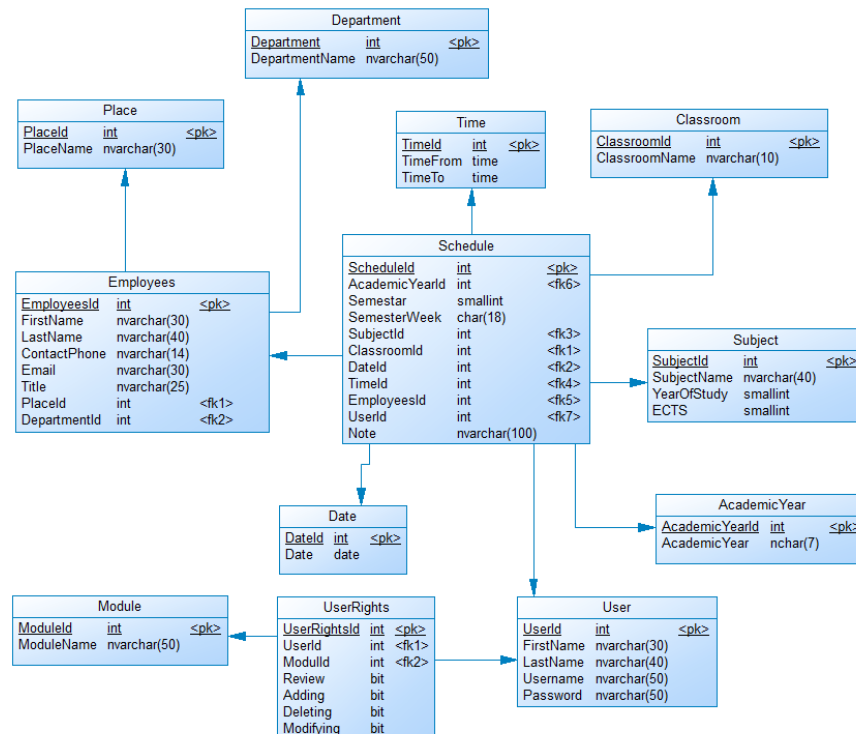


Figure 3. Physical model of the database

The proposed solution was developed based on user requirements. In their case, the schedule is created for each upcoming week at the end of the previous week and is generally not the same every

week. For example, lectures or exercises for a specific subject do not take place every week. Additionally, in weeks when they do occur, they may not happen on the same day or at the same time. Therefore, a Date table is essential.

#### 4. Development of the Application Middle Layer

To develop the application middle layer, the Java programming language was used in the Eclipse IDE for Java Developers – 2023-12. At the beginning of the development of the middle layer, it is necessary to establish a connection to the database.

After launching the application, links must be visible that lead to the registration and login of the department secretary or administrator, as well as to the viewing of the class schedule by study program, classroom, or professor.

The method for registering a new user is implemented using the SQL INSERT command. The functionality of the login form is based on a simple principle: executing a SELECT query on the Users table based on the data entered by the user. If the code and username match, the user is logged into their account. After logging in, there are options for searching the schedule by study program, classroom, or professor, as well as adding, deleting, or updating information about lectures and course coverage. Other forms should have the functionality for viewing, adding, modifying, and deleting data while maintaining the referential integrity of the database.

It is important to note that the class schedule for the following week is locked during certain time slots in the previous week, meaning that secretaries cannot add, delete, or update it. The administrator has the ability to change the locking periods of the schedule. In addition to this, the administrator can add, delete, and update information about professors, departments, classrooms, and subjects while maintaining the referential integrity of the database.

#### 5. User Interface Design

In the development of the software solution, the next step is the design of the user interface. To create the user interface, a set of components was used, including components for standard HTML elements, which are part of the Vaadin Design System. These components include various input fields (such as TextField, DatePicker, TimePicker, EmailField, PasswordField, etc.), buttons, menu bars, tables, side navigation, and layout options (FormLayout, VerticalLayout, HorizontalLayout, and others).

The image displays two screenshots of a web application interface for a 'Faculty of Philosophy Class Schedule' on a mobile phone. Both screenshots show a form with filters for 'Study program' (set to 'Philosophy'), 'Year of study' (set to '1'), and 'Week' (set to '14.10.-19.10.2024').

The left screenshot shows a table with columns: Time, Monday, Tuesday, Wednesday, and Thursday. The table lists several lecture slots, including 'History of Philosophy Lectures, Classroom: 105' and 'Democracy and Human Rights Lectures, Classroom: 105'.

The right screenshot shows a zoomed-in view of the Monday and Tuesday columns. It highlights a specific lecture slot on Monday at 08:45-09:30, 'History of Philosophy Lectures, Classroom: 105'. Below this, there is a dropdown menu for 'Democracy and Human Rights' and a 'Lectures' dropdown. At the bottom, there are 'Next' and 'Copy to Next' buttons.

Figure 4. Class schedule web forms on a mobile phone



The previous image (Figure 4) shows the web form of the class schedule on a mobile phone in preview mode and in input mode.

Figure 5 shows an example of a web form for the schedule of classes at the Faculty of Philosophy for one week on a laptop.

**Faculty of Philosophy**

**Class Schedule**

Study program: Philosophy Year of study: 1 Week: 14.10.-19.10.2024

Time	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
08:00 - 08:45	History of Philosophy Lectures, Classroom: 105			English Language Lectures, Classroom: 204		
08:45 - 09:30	History of Philosophy Lectures, Classroom: 105			English Language Lectures, Classroom: 204	Psychology Exercises, Classroom: 107	
09:45 - 10:30	Democracy and Human Rights Lectures, Classroom: 105	Psychology Lectures, Classroom: 107	Philosophical Techniques Lectures, Classroom: 110	English Language Exercises, Classroom: 204	History of Philosophy Exercises, Classroom: 105	
10:30 - 11:15	Democracy and Human Rights Lectures, Classroom: 107	Psychology Lectures, Classroom: 107	Philosophical Techniques Lectures, Classroom: 110	English Language Exercises, Classroom: 204	Philosophical Techniques Exercises, Classroom: 110	
11:30 - 12:15	Democracy and Human Rights Lectures, Classroom: 107	Introduction to Philosophy Lectures, Classroom: 108	Introduction to Philosophy Exercises, Classroom: 108		Philosophical Techniques Exercises, Classroom: 110	
12:15 - 13:00	Democracy and Human Rights Lectures, Classroom: 107	Introduction to Philosophy Lectures, Classroom: 108	Introduction to Philosophy Exercises, Classroom: 108			
13:15 - 14:00		Democracy and Human Rights Exercises, Classroom: 107				
14:00 - 14:45		Democracy and Human Rights Exercises, Classroom: 107				
	<span>Edit</span>	<span>Edit</span>	<span>Edit</span>	<span>Edit</span>	<span>Edit</span>	<span>Add</span>

Figure 5. Class Schedule form

## 6. Application Testing

In this phase, application testing was conducted. First, unit testing was performed, focusing on individual functionalities. After successful unit testing, integration testing was carried out, which involves testing the interaction between the user interface layer and the middleware, as well as between the middleware and the data layer. At the end of this process, user testing was conducted to gather feedback from users, leading to minimal adjustments to the application. After testing, documentation for users (help) was prepared to assist them in using the application, along with technical documentation for the development team related to the data model, code structure, and application architecture.

## 7. Conclusion

The developed application offers many possibilities for employees in creating class schedules for lectures and exercises. It is designed to be used by faculties and various schools. All employees and students can access the lecture schedule via their mobile phones, as the application meets all responsive design requirements. The proposed approach emerged as a result of a project task in the second cycle of studies at the Faculty of Philosophy, University of East Sarajevo, in the Mathematics and Computer Science program. The advantages of the proposed approach are as follows: a responsive application, the use of the latest technologies for development, the ability to apply it in cases where the schedule is not the same every week, and the capability for quick creation for each day and each week separately.

Like any application, this one has the potential for further development. One suggestion for future enhancement is to create a communication feature that allows students to interact with the staff responsible for creating the schedules. This communication would involve students providing suggestions regarding the timing of specific classes.

## References:

- [1] D. Mijić, Uvod u veb programiranje, Elektrotehnički fakultet, Istočno Sarajevo, Akademska misao, Beograd, 2019.

- [2] M. Dobrojević, N. Baćanin-Džakula, Veb programiranje, Prvo izdanje, Univerzitet Singidunum, Beograd, 2021.
- [3] [https://www.ibm.com/topics/three-tier-architecture\\_](https://www.ibm.com/topics/three-tier-architecture_) (date of access: 22. 01. 2024).
- [4] D. Krneta, Baze podataka, My Book, Banja Luka, 2023.
- [5] <https://vaadin.com/faq> (date of access: 23. 01. 2024).
- [6] J.J. Laguardia, J.A. Flores, University Class Schedule Assignment by a Tabu Search Algorithm, 2022 8th International Engineering, Sciences and Technology Conference (IESTEC), Panama, Panama, 2022, pp. 728-732.
- [7] M.K. Kakkar et al, Class Schedule Generation using Evolutionary Algorithms, 2021 J. Phys.: Conf. Ser. 1950 DOI: 10.1088/1742-6596/1950/1/012067.
- [8] K.K.L. Ho and M.Lu, Web-based expert system for class schedule planning using JESS, IRI -2005 IEEE International Conference on Information Reuse and Integration, Conf, 2005., Las Vegas, NV, USA, 2005, pp. 166-171.
- [9] P.F. Oliveira, P. Matos, Mobile Integration for Sakai LMS and Class Schedule Platform at an Higher Education Institution, 2023 International Conference on Engineering and Emerging Technologies (ICEET), Istanbul, Turkiye, 2023, pp. 1-6.
- [10] R. Ismibayli, S. Rzayeva, An Interactive Intelligent System of Creating a Class Schedule. In: G. Mammadova, , T. Aliev, K. Aida-zade, (eds) Information Technologies and Their Applications. ITTA 2024. Communications in Computer and Information Science, vol 2225. Springer.
- [11] V. Tomašević, Razvoj aplikativnog softvera, Univerzitet Singidunum, Beograd, 2017.
- [12] N. Frid, A. Jović, Modeliranje programske potpore UML – dijagramima, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, 2023.
- [13] M. Veinović, G. Šimić, Uvod u baze podataka, Univerzitet Singidunum, Beograd, 2010;

# Experience with Proof-of-Concept Applications with a Micro-frontend Architecture: The Vet Clinic Box Monitoring App Case

Katarína Semanová<sup>1</sup>, Csaba Szabó<sup>2</sup>

<sup>1</sup> Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovakia

<sup>2</sup> Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovakia

katarina.semanova@student.tuke.sk; csaba.szabo@tuke.sk

## Abstract:

Many times, architectural decisions determine the future of an application. Among different distributed architectures, microservices and micro-frontends aim to support effort distribution between a larger number of development teams by separating functionality. This paper presents evaluation of the proof-of-concept Vet Clinic Box Monitoring App, which utilizes a micro-frontend architecture. It discusses the design of the system and its components, focusing on modularity and scalability. Subsequently, it details the testing and evaluation procedures employed to assess the performance and reliability of the application. Our findings indicate the efficacy of micro-frontend architecture in enhancing development speed and overall user experience. Furthermore, this architecture supports system extension in an arbitrary way, such as inclusion of artificial intelligence applications for domain specific tasks. In our case, it led to a new concept of a Smart Pet Care Application.

## Keywords:

Micro-frontend architecture, vet clinic box monitoring, web application

## 1. Introduction

The rise of micro-frontend architecture presents a promising solution to tackle the complexities inherent in constructing sophisticated web applications. The segmentation of extensive monolithic systems into smaller, self-contained frontend modules is a key aspect of micro-frontend architecture. By breaking down complex systems into manageable parts, developers can focus on individual components more effectively, leading to improved team independence. Each frontend module operates independently, allowing teams to work on different parts of the application without being constrained by dependencies or conflicts with other modules. This modular approach enhances agility and flexibility in development, enabling teams to iterate rapidly and respond quickly to changing requirements. Accelerated development timelines, and elevated user satisfaction are another significant benefit of this architecture. This paper presents an assessment of our proof-of-concept Vet Clinic Box Monitoring App, a web-centric platform designed to streamline veterinary clinic management, all crafted within the framework of micro-frontend architecture. It serves as a compelling example of how micro-frontend architecture can be used in modern web development practices.

Our architectural decision was already proven during the evaluation of the proof-of-concept, the partner vet clinic presented an unexpectedly high customer appetite for features during the requirement evaluation period.

## 2. Micro-frontend architecture

Micro-frontend architecture is a concept in web application architecture that focuses on dividing the frontend into multiple smaller and independent parts. This approach allows developers to work on different parts of the web application independently, thereby increasing development speed and enhancing the application's scalability. The architecture leverages frameworks that assist in the

development of micro-frontend-based applications. Figure 1 illustrates a micro-frontend application, where each part is its own frontend with its own build and testing phases and can be deployed to production independently of other parts of the application, as described in [1, 2]. For creating micro-frontends, medium-to-large projects [3] with different teams are most suitable as it helps streamline the development process. Micro-frontends are often used for web projects since existing functionalities cannot be changed in native applications for iOS and Android, as mentioned in source [4].

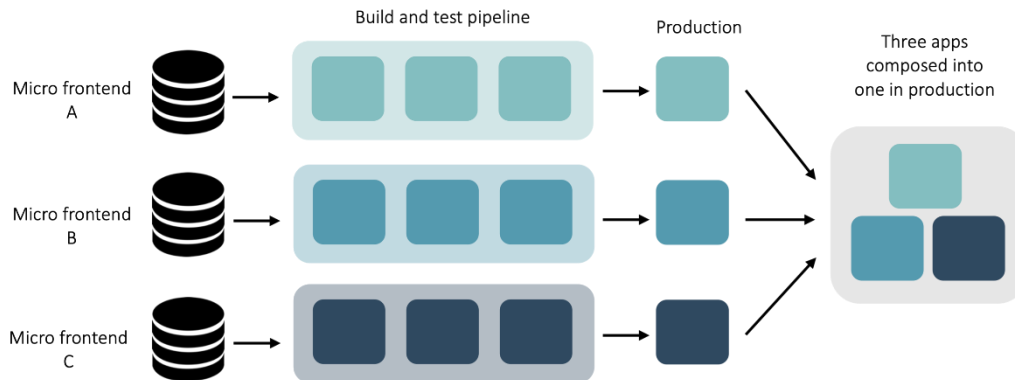


Figure 1: Schema of the micro-frontend architecture. [5]

## 2.1. Advantages of micro-frontend architecture

Micro-frontend architecture offers a development approach [6] that is not tied to a specific technological stack, providing flexibility in design and overall development. Its characteristic advantages include [7]:

- **Code Separation:** Micro-frontend frameworks facilitate code management and updates by dividing the code into smaller parts, making the entire project more manageable.
- **Innovation:** Microservices and micro-frontend architecture allow for technology agnosticism, fostering innovation in technology selection.
- **Fault Isolation:** Micro-frontend design provides an advantage in creating extensive applications. If a failure occurs in one of the micro applications [8], it does not pose any risk to the entire system.
- **Scalability:** This architecture allows for adjusting the size of individual components of the system, for the necessary scaling up or down of the overall capacity of the application [9]. Application parts can be changed without affecting the overall speed of the online application.

## 2.2. Disadvantages of micro-frontend architecture

Unorganized use of micro-frontends can lead to unnecessary complexity as it requires understanding dependencies among different modules. When using them, it is also necessary to reconsider some negative aspects, such as [10,11]:

- **Operational Complexity:** Higher costs and resource consumption are caused by separate deployment of development environments and servers when using micro-frontends.
- **Overhead:** Necessary programs may cause issues for users with less powerful infrastructure.
- **Domino Effect:** Due to the interdependency of technologies, the occurrence of an error in any technology may negatively impact overall performance.

## 3. Comparison of software architecture

For the utilization of multiple sensors simultaneously, among the three software architectures depicted in Table 1 below, micro-frontend architecture appears to be the most suitable. It can be observed that micro-frontends are like microservices but is more focused on the frontend part, and thus,

on the user interface. Each sensor will have its own micro-frontend, meaning that each will have its own source code and database. This division allows for better scalability, extensibility, and collaboration.

**Table 1:**

Comparison of monolithic, microservices and micro-frontends software architectures.

Characteristic	Monolith	Microservices	Micro-frontends
Architecture	Single application	Service-oriented architecture	Frontend division
Scope	One large application	Independent services	Small components for each sensor
Scalability	Vertical	Horizontal	Horizontal
Development and maintenance	More complex	Less complex	Simple
Dependencies	Strong dependencies between different parts of the application	Weak dependencies between services	Minimal or no dependencies between components
Technologies	Uniform technology for the entire application	Different technologies for each service	Flexibility in choosing technologies for each micro-frontend
Security	One large application requires a comprehensive security approach	Security must be ensured at the level of each service	Simpler security implementation for each micro-frontend

#### 4. Communication between micro-frontends

The communication layer between micro-frontends largely depends on the type of composition chosen. According to the article [12], the goal of communication is to inform other micro-frontends about user interaction. To ensure that different teams are loosely coupled and to leverage independent deployment, communication must work at a level where individual interfaces, frontends, are not aware of the existence of others.

Figure 2 describes various types of communication in micro-frontend architecture [13, 14]. The first point in the figure illustrates ways of communication within the frontend part of the application in a web browser. The second point illustrates how the frontend retrieves data from the backend part of the application. The last point demonstrates data replication between teams.

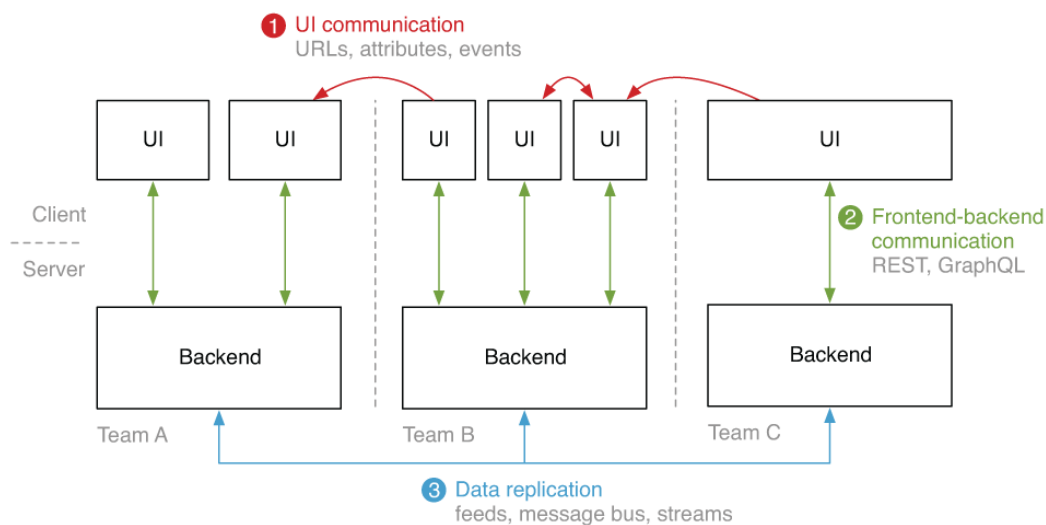


Figure 2: Communication between micro-frontends.

## 5. Vet Clinic case study

The Vet Clinic Box Monitoring App consists of several distinct frontend modules, each tasked with handling specific sensor functionalities. These functionalities include measuring air temperature, humidity, noise levels, weight, and camera monitoring. These modules are developed and deployed separately, allowing teams to work autonomously and integrate their components seamlessly into the application. The micro-frontend architecture enables dynamic loading of modules at runtime, resulting in faster page rendering and improved user interaction.

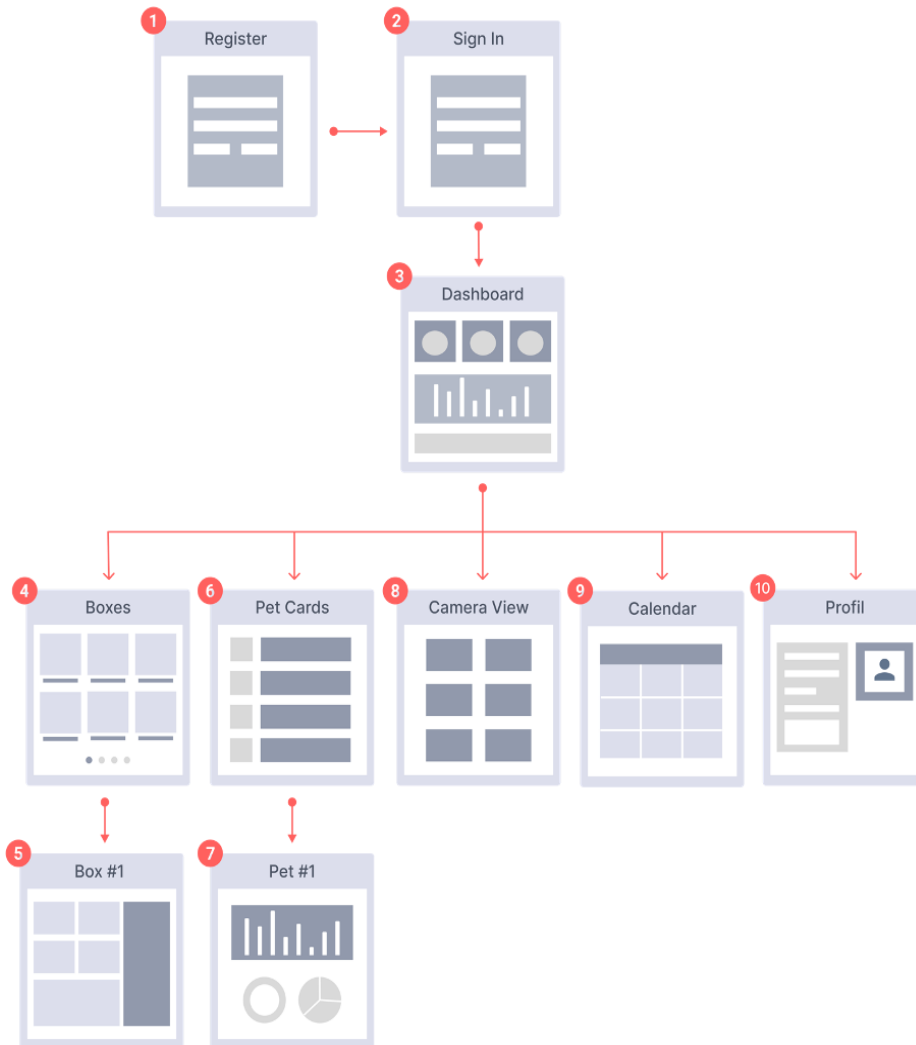


Figure 3: Application screens.

Figure 3 is aimed at concluding the key features as views/screens identified during requirements elicitation. The system should provide features related to user identification (1,2,10), an overview of monitoring results (3), vet clinic box management (4,5), pet identification and management (6,7), real-time camera stream connections (8) and time organization (9).

The following subsections introduce the primary micro-applications within this web application, providing descriptions of their functionalities and implementations.

### 5.1. Root-app

Root-app serves as the fundamental micro-application, functioning as the primary entry point for launching the entire application. It is designed to provide an intuitive interface among the various sensor components of the application, incorporating essential navigation elements to facilitate integration with

other micro-applications. Additionally, Root-app offers a login interface allowing users to create accounts or log in using their Google accounts, implemented through the Firebase service. This component of the application also facilitates data monitoring generated by the sensors. Root-app plays a pivotal role in facilitating communication among all micro-applications.

## 5.2. Root-config

Root-config is a micro-application responsible for configuring the application. It contains a configuration file listing all micro-applications to be loaded. This file is utilized during application startup for dynamically loading and initializing individual sensor micro-applications.

## 5.3. Sensors

Sensor micro-applications, such as Temperature, Weight, and Humidity, represent standalone entities. Each of these applications is written in TypeScript and utilizes the React framework. Individual micro-applications handle the display and subsequent processing of data from specific sensors. For each sensor, a separate micro-application is designed, which is independent and has minimal dependencies on other micro-applications. This enables easy addition of sensors to the system without significant modifications to existing parts of the application. Real physical sensors are not used in the implementation; instead, they are simulated with mock data for the purposes of this proof-of-concept.

There were numerous challenges identified related to the sensors, such as their required independence from pets. This contributes to a further improvement in the sensor implementation and variety.

## 6. Evaluation and testing

To ensure the proper functioning of the Vet Clinic Box Monitoring App, testing is essential, forming an integral part of software development to guarantee its quality, reliability, and efficiency. For the testing of this web application, automated testing has been chosen, primarily focusing on end-to-end tests using the Jest tool. This approach allows for repeated testing of the application without the need for manual intervention. In this way, it ensures that the application will continue to function correctly even after the addition of new features or potential updates.

The results of automated testing have confirmed that the application functions correctly and is ready for deployment and subsequent use. These are evidenced by the following testing outcomes:

1. **Comprehensive Testing:** All major components of the application underwent thorough testing using various test scenarios covering essential functionalities. This ensured that all features are subjected to testing.
2. **No Errors or Deficiencies:** The absence of errors or deficiencies throughout the testing process represents success. This verifies that the application is stable and reliable, which is a key factor in ensuring a positive user experience.
3. **Requirement Fulfillment:** The test scenarios successfully confirmed that the application meets the requirements set forth in the initial development phase. This fact confirms that the application is ready for deployment and usage.

Principles of micro-frontend architecture were used to integrate multiple micro-frontends into a single web application. This approach allows for the creation of a unified application comprising several independent micro-applications.

Figure 4 illustrates the result of such integration, where highlighted sections represent individual micro-frontends. Each of these micro-frontends is responsible for displaying data from a single sensor located in an animal box. The remaining portion of the screen displays the main part of the application, i.e., the root-app, which includes a navigation menu and basic information about the animal patient and its owner.

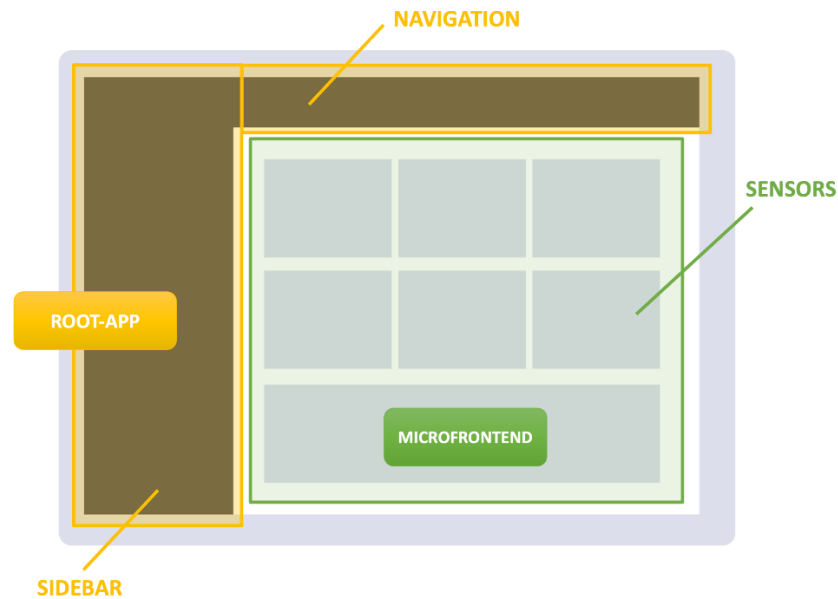
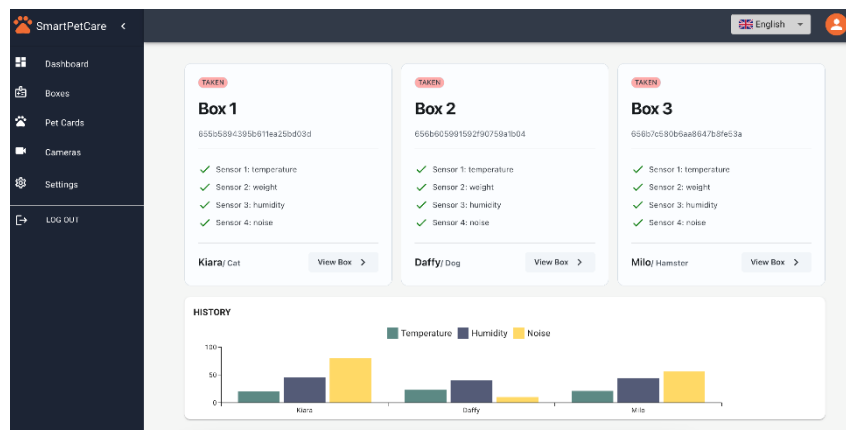
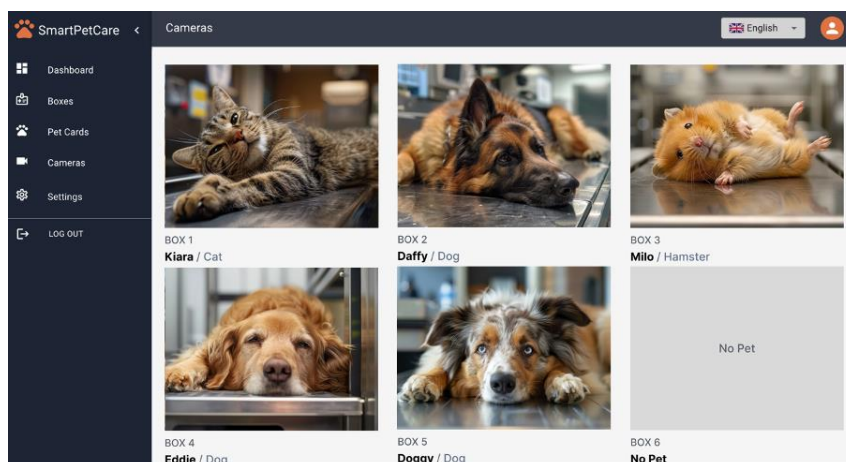


Figure 4: Application division into micro-frontends.

Figure 5 includes selected screenshots of the Smart Pet Care application, which is aimed at being the final product of our research and development.



a) Dashboard



b) Camera view

Figure 5: Application screenshots.



One of the key benefits of micro-frontend architecture is its enhanced maintainability. It evaluated the ease of maintenance by examining the process of updating individual modules, adding new features, and fixing bugs. The modular nature of the architecture allowed for independent development, reducing the risk of introducing regressions and conflicts. Furthermore, versioning and dependency management mechanisms facilitated seamless integration of updates across modules.

## 7. Conclusions

In conclusion, our systematic evaluation of the Vet Clinic Box Monitoring App demonstrates the effectiveness of the micro-frontend architecture in building scalable, reliable, and maintainable web applications. The modular design enables teams to work autonomously, accelerates development cycles, and improves overall user experience. The successful implementation of micro-frontend architecture in this case study underscores its potential in web application development, particularly in domains like veterinary clinic management where real-time monitoring and seamless integration of sensor data are critical.

The assessment revealed several key insights that underscore the effectiveness of micro-frontend architecture in building scalable, reliable, and maintainable web applications. The modular nature of micro-frontend architecture significantly enhances maintainability. By breaking down the frontend into independent micro-applications, development becomes more manageable. This modularity reduces the risk of conflicts and regressions, as updates can be implemented independently. Versioning and dependency management mechanisms further ensure seamless integration of updates, simplifying maintenance tasks and promoting stability over time.

Moreover, micro-frontend architecture offers unparalleled scalability and flexibility. This architecture allows for the easy addition of new sensors or functionalities without extensive modifications to existing components. Each micro-frontend operates autonomously, enabling teams to work on different parts of the application independently. This flexibility is crucial for accommodating future expansion and adapting to evolving user needs, ensuring that the application remains agile and responsive to changing requirements.

Additionally, the comprehensive testing procedures confirmed stability, reliability, and adherence to requirements of the Vet Clinic Box Monitoring App. Automated end-to-end testing using tools like Jest ensured thorough validation of essential functionalities. These positive testing outcomes instill confidence in the deployment and usage of the application, affirming its readiness for real-world usage. While the study provides compelling evidence of the benefits of micro-frontend architecture, there remain avenues for future research and optimization. Further exploration could focus on refining communication patterns between micro-frontends, optimizing resource utilization to mitigate operational complexities, and investigating techniques for enhancing security in distributed frontend architectures.

In summary, the Vet Clinic Box Monitoring App exemplifies the transformative potential of micro-frontend architecture in web application development. By embracing modularity, scalability, and autonomy, micro-frontend-based solutions offer a promising paradigm for building next-generation web applications that deliver superior user experiences and adaptability to changing requirements.

During the requirement engineering process, we faced common domain challenges. Later, the so-called customers' appetite rose unexpectedly fast, which turned out to be the proof of the initial architectural decision using micro-frontend architecture in our solution. There is still work in progress to implement artificial intelligence techniques and models [15] to achieve a Smart Pet Care App. Because of the right decision, all new system components can be either developed independently or reused from independent systems after wrapping them to fit our component structure.

### Acknowledgment:

The authors would like to thank the staff of the Clinic of Birds, Exotic and Free-Living Animals of University Veterinary Hospital at the University of Veterinary Medicine and Pharmacy in Košice for providing domain-specific information and requirements for development of the referred proof of concept application.

## References:

- [1] H. Zhang, "Understanding the differences between event-driven, message-driven, and microservices architectures with AWS services," Feb 2023. [Online]. Available: <https://faun.pub/what-is-difference-of-event-driven-architecture-message-driven-architecture-and-microservices-f5623e51f868>
- [2] G. Liu, B. Huang, Z. Liang, M. Qin, H. Zhou and Z. Li, "Microservices: architecture, container, and challenges," *2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, Macau, China, 2020, pp. 629-635, doi: 10.1109/QRS-C51114.2020.00107.
- [3] I. Nadareishvili, R. Mitra, M. McLarty, and M. Amundsen, "Microservice Architecture: Aligning Principles, Practices, and Culture" (1st. ed., 2016). O'Reilly Media, Inc.
- [4] M. Techlabs, "What is micro-frontend? Benefits of using micro-frontend architecture," Nov 2022. [Online]. Available: <https://medium.com/geekculture/what-is-micro-frontend-benefits-of-using-micro-frontend-architecture-f6d667edb03d>
- [5] Adservio, "The Dark Side of Micro Frontends" Feb 2023. [Online]. Available: <https://www.adservio.fr/post/the-dark-side-of-micro-frontends#el6>
- [6] K.J. Hole, "Tutorial on systems with antifragility to downtime," *Computing* **104**, 73–93 (2022). <https://doi.org/10.1007/s00607-020-00895-6>
- [7] A. Bandeira, C. A. Medeiros, M. Paixao and P. H. Maia, "We Need to Talk About Microservices: an Analysis from the Discussions on StackOverflow," *2019 IEEE/ACM 16th International Conference on Mining Software Repositories*
- [8] J. Sorgalla, F. Rademacher, S. Sachweh, and A. Zündorf, "Collaborative model-driven software engineering and microservice architecture: a perfect match?" In *Proceedings of the 19th International Conference on Agile Software Development: Companion (XP '18)*. Association for Computing Machinery, New York, NY, USA, Article 24, 1–2. <https://doi.org/10.1145/3234152.3234194>
- [9] N. Tanković, T. Galinac-Grbac, "Run-time interpretation of information system application models in mobile cloud environments," *Computer Science and Information Systems 2020* Volume 17, Issue 1, Pages: 1-27, <https://doi.org/10.2298/CSIS180829021T>
- [10] A. Gromenko, "Should your projects have micro-frontends? — code-care," Oct 2022. [Online]. Available: <https://code-care.com/blog/should-your-2022-projects-have-micro-frontends/>
- [11] M. Samardžić, R. Šajina, N. Tanković and T. G. Grbac, "Microservice Performance Degradation Correlation," *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, Opatija, Croatia, 2020, pp. 1623-1626, doi: 10.23919/MIPRO48935.2020.9245234.
- [12] D. Minkovski, "Micro-frontends and their benefits - better programming," May 2022. [Online]. Available: <https://betterprogramming.pub/micro-frontends-and-their-benefits-674cce350ecc>
- [13] Manning.com, "6 communication patterns," 2023. [Online]. Available: <https://livebook.manning.com/book/micro-frontends-in-action/chapter-6/>
- [14] Y. Abgaz *et al.*, "Decomposition of Monolith Applications Into Microservices Architectures: A Systematic Review," in *IEEE Transactions on Software Engineering*, vol. 49, no. 8, pp. 4213-4242, Aug. 2023, doi: 10.1109/TSE.2023.3287297.
- [15] M.A. Langford, S. Zilberman, and B. Cheng, "Anunnaki: A Modular Framework for Developing Trusted Artificial Intelligence," *ACM Trans. Auton. Adapt. Syst.* **19**, 3, Article 17 (September 2024), 34 pages. <https://doi.org/10.1145/3649453>

# The Perception of Risks in Digital Payment Services

Nikola Jerković <sup>1</sup>, Ana Jovanović <sup>1</sup>, Ivana Berković <sup>1</sup>, Edit Boral <sup>2</sup>

<sup>1</sup> Technical Faculty "Mihajlo Pupin" Djure Djakovica bb, Zrenjanin, Republic of Serbia

<sup>2</sup> Johns Hopkins University Center for Talented Youth, New York, USA

nikolajerkovic13@gmail.com; jovanovica129@gmail.com; ivana.berkovic62@gmail.com;  
edit.boral@gmail.com

## Abstract:

The development of digital payment services has significantly transformed the way everyday transactions are conducted, offering greater flexibility, speed, and convenience. Over time, digital payments have become an essential tool in modern business, pushing physical cash to the background due to the numerous advantages they provide. However, as these services become more widespread, the number of potential risks that users face is also increasing. Threats such as fraud, data theft, and unauthorized access pose serious challenges to the security of digital payment systems. This paper aims to assess the level of trust in the security of digital payments and explore the presence of risks and fraud within these systems.

## Keywords:

digital payment services, risks, security

## 1. Introduction

The retail sector has undergone significant evolution towards digitalization in recent decades. This process began with the advent of the Internet, which enabled access to global information and connected consumers to various markets. Over time, the integration of innovative digital technologies, business models, and platforms has become crucial for the development of a multilayered market. These changes have shaped the way people make purchases and, consequently, their consumer habits, leading to the widespread adoption of digital services [1].

Digital payment, also known as electronic payment, is a modern method of conducting transactions through digital platforms. This form of payment is characterized by the absence of physical cash, and instead, digital channels are used to send and receive money [2], [3]. All transactions in digital payment are carried out via the Internet, making this method fast and convenient for users. To perform digital payments, both the payer and the recipient must have a bank account and access to appropriate digital devices such as mobile phones, computers, or payment cards [3].

### 1.1. Risks in digital payments services

With the increasing use of digital commerce, users face a range of risks while browsing online stores and making decisions about electronic payments [1]. Some of the most common risks in digital payment and commerce include mobile malware, phishing, physical theft, data extraction, duplicating SIM card and credit card fraud [4].

- Mobile phone viruses attack private messages and operating systems [4], and they most commonly spread via Bluetooth and MMS [5]. While Bluetooth viruses can affect a wide base of vulnerable users, their spread is limited by the short range and the dynamic movement of people. Unlike Bluetooth viruses, MMS viruses use a topological spreading approach that does not depend on user movement, allowing the virus to quickly infect a large number of devices. However, their spread is limited by the presence of different operating systems on devices, which divides the user network and makes it harder for the virus to reach all devices [5]. Viruses can lead to the theft of sensitive data and financial losses [4].

- Phishing attacks are used to deceive users into revealing confidential information, such as banking details, often through fake websites, emails, or SMS messages [4]. Attackers continuously refine their methods, creating fake websites that resemble legitimate ones, which further increases the risk of fraud. Phishing often leverages publicly available personal information to target specific victims, particularly in spear-phishing attacks, where the attacks are based on data such as banking details or recent online purchases. Many phishing attacks also involve viruses like worms or trojans, which compromise the security of devices and enable further attacks [6].
- Physical theft poses a serious risk as it allows thieves access to critical devices and information related to digital payments, such as bank cards, identification documents, or mobile phones used for transactions [4]. This type of threat can occur through hardware theft, vandalism, or unauthorized access, which can result in the loss of sensitive data [7].
- Data extraction is a threat that involves using bots to collect sensitive information from websites, often for the purpose of resale or misuse. The process includes identifying targeted URLs and parameters, deploying data extraction tools that can overload servers and reduce performance, as well as gathering content and databases for further analysis and potential abuse [8].
- SIM card cloning allows fraudsters to duplicate the functionality of the original card [4]. Older SIM cards, especially those with weaker security algorithms, are particularly vulnerable to these attacks [9]. However, despite modern SIM cards having self-destruct mechanisms that limit the number of decryption attempts, attackers continually develop more sophisticated methods for cloning [9]. The consequences of SIM card cloning include financial fraud, such as access to bank accounts and unauthorized transactions, which can lead to monetary losses. Personal data from the cloned SIM card can be hacked, and attackers can use the cloned card to commit fraud and open credit in the user's name. Additionally, the user's contacts may receive suspicious messages, leading to further issues [10].
- Credit card fraud occurs when fraudsters use someone else's financial information to issue new cards or take control of existing accounts. Fraud can also happen through the theft of a new credit card before it reaches its owner. The use of a lost or stolen credit card for illegal purposes is another common form of fraud. Additionally, fraud can arise when fraudsters use unlawfully obtained personal information and bank accounts to gain control of a bank account. In this way, fraudsters impersonate the true cardholders, requesting mail to be redirected to a new address or reporting the card as lost. Furthermore, the creation of fake and counterfeit cards poses a significant threat [11].

## 1.2. Risk perception

Previous studies by authors such as Jacoby, Kaplan and Mitchell have indicated that risks significantly influence users' decisions regarding the use of digital payment services. Risk perception in digital payments is a complex and dynamic process that depends on numerous factors. A decrease in risk perception directly increases users' trust in digital payment services. Trust, as Costa points out, is linked to vulnerability, expectations, and individual control, making it a key factor in decision-making related to digital payment services. Trust in institutions providing digital payment services also plays a significant role. The credibility of such companies can be established in various ways, which enhances consumer trust in their security during digital transactions. However, insufficient awareness of the risks can lead to the avoidance of these services. Additionally, concerns about personal data frequently cause users to abandon online transactions. Furthermore, previous experiences have a considerable impact on users' risk perception. Positive experiences can increase trust in digital services, while negative experiences often lead to caution and doubt regarding future transactions. In this regard, it is certainly important to take appropriate security measures to ensure that users feel safe when making digital payments [5].

## 2. Research methodology

The research addresses the challenges of digital payments in Serbia, with a particular focus on how users perceive risks related to digital payment services.

The following presents the survey questions from this research, along with the respondents' answers, in which they expressed their views on the risks of digital payments. This research involved 100 participants who are users of digital payment services.

For the first question, "How often do you use digital payment services?", 31% of respondents indicated that they use digital payment services several times a month, 27% use them several times a week, 23% use them rarely, and 19% use them everyday.

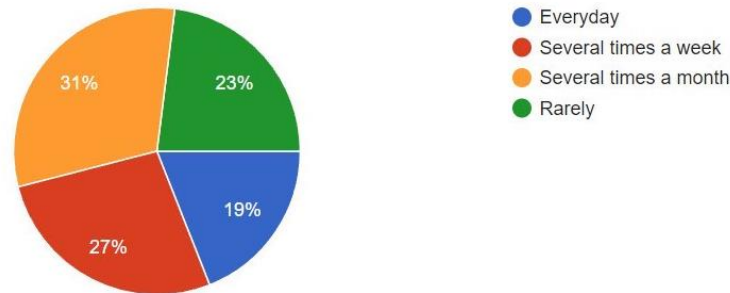


Figure 1: Answer the question "How often do you use digital payment services?"

The second question, "How secure do you feel when making payments via digital payment services?" offers a rating scale from 1 to 5, where 1 represents "not at all" and 5 represents "very secure". The largest number of respondents, 39%, rated their sense of security as 4. The next largest group, 31%, gave a rating of 5. 18% of respondents rated their security level as 3, while 9% gave a rating of 2, and only 3% rated their security as 1.

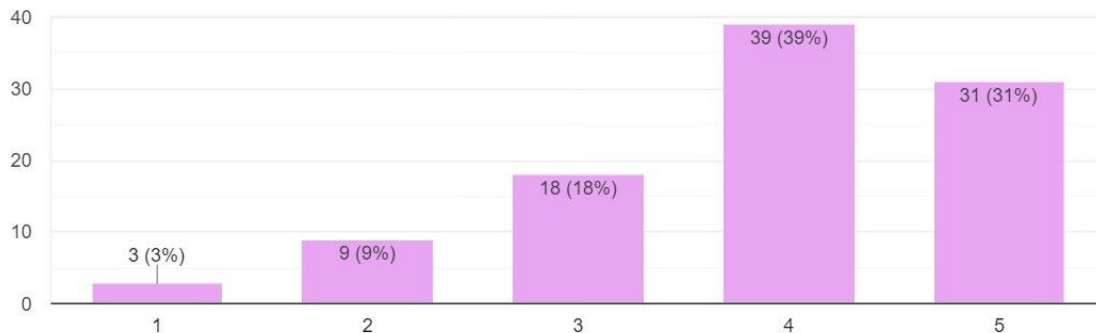


Figure 2: Answer the question "How secure do you feel when making payments via digital payment services?"

In response to the question, "How familiar are you with the various threats and risks associated with digital payments?", 66% of respondents stated that they are partially familiar with the threats and risks in digital payments, while 29% reported being completely familiar with these risks. Only 5% of respondents indicated that they are not at all familiar with the risks.

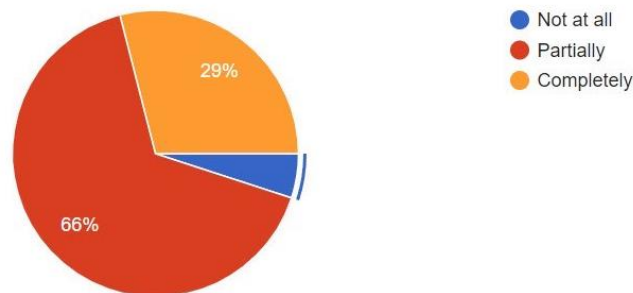


Figure 3: Answer the question "How familiar are you with the various threats and risks associated with digital payments?"

Based on the responses to the question, "Have you ever been a victim of any of the following risks in digital payments: mobile malware, phishing, physical theft, data extraction, duplicating SIM card or credit card fraud?", it can be concluded that as many as 90% of respondents have never been a victim of any of these risks, while the remaining 10% reported having directly encountered some risk.

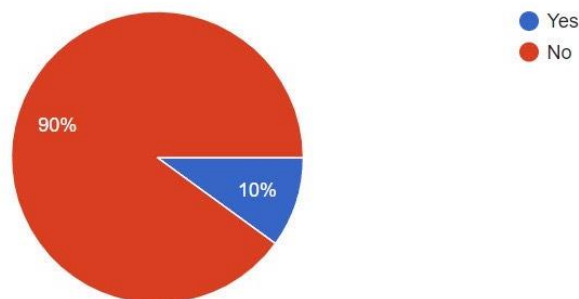


Figure 4: Answer the question "Have you ever been a victim of any of the following risks in digital payments: mobile malware, phishing, physical theft, data extraction, duplicating SIM card or credit card fraud?"

In response to the question, "Have you ever avoided online shopping or payments due to concerns about the security of your data?", 68% of respondents answered affirmatively, indicating that they have indeed avoided online shopping or payments due to security concerns, while 32% have never avoided online shopping or payments for such reasons.

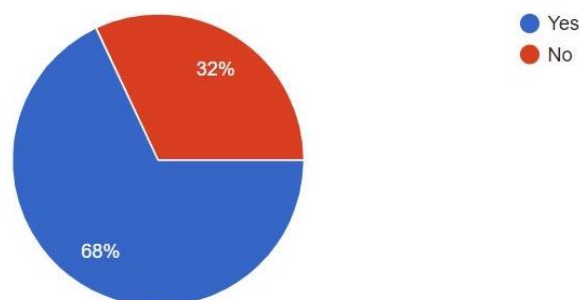


Figure 5: Answer the question "Have you ever avoided online shopping or payments due to concerns about the security of your data?"

In response to the question, "How often do you take certain steps to protect your own data in digital payment systems (using strong passwords and PINs, avoiding suspicious apps and links, regularly

updating, avoiding public Wi-Fi networks, etc.)?", the answers are as follows: 52% of respondents answered "very often," 42% responded "rarely," and 6% said "never."

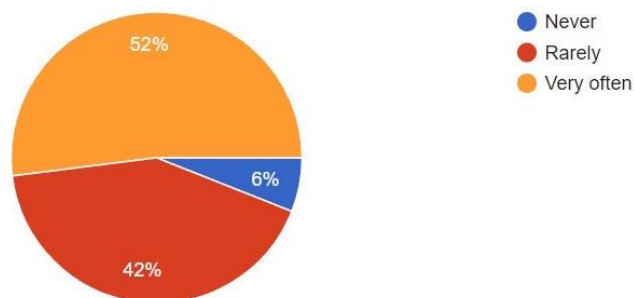


Figure 6: Answer the question "How often do you take certain steps to protect your own data in digital payment systems (using strong passwords and PINs, avoiding suspicious apps and links, regularly updating, avoiding public Wi-Fi networks, etc.)?"

For the final question in the survey, "How much trust do you have in institutions that provide digital payment and transaction services?", 39% indicated they have partial trust, 37% reported having quite a lot of trust, 14% expressed only a little trust, 9% said they trust these institutions completely, while 1% stated they do not trust them at all.

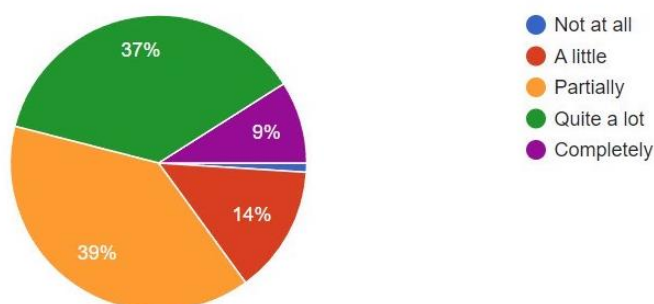


Figure 7: Answer the question "How much trust do you have in institutions that provide digital payment and transaction services?"

### 3. Comparison with Visa's study

The survey "Plaćaj bezbedno" conducted by Visa in 2023 revealed that Serbian citizens exhibit a high level of caution regarding online payment security. The analysis showed that they are among the most cautious in the region concerning fraud in digital transactions. Supporting this is the fact that only 28% of respondents had encountered any form of risk in this area, the lowest percentage compared to other countries surveyed. Over half of the participants (52%) believe they can identify threats and risks in digital payment services. The study also found that users in Serbia are highly skeptical of messages related to financial winnings, package statuses, free gifts, or password requests. However, they tend to overlook key indicators of fraud, such as suspicious email addresses and grammatical errors, which are significant warning signs [12].

The conducted research described in the paper revealed certain similarities with Visa's study regarding the perception of risks associated with digital transactions. Like in the Visa study, respondents exhibited a high level of caution, with as many as 68% admitting to having avoided online shopping or payments at least once due to security concerns. This caution is accompanied by a feeling of insecurity, as most respondents, while generally confident, admitted that they do not feel entirely secure when conducting digital transactions. Furthermore, respondents' trust in institutions providing digital payment services shows a certain degree of reservation, with most indicating partial trust in these institutions.

Additionally, 52% of respondents in this study often take security measures to protect their personal data within digital payment systems. The respondents also emphasized that they are generally at least partially aware of the risks associated with digital payment services, with 29% stating they are fully aware of such risks. Accordingly, only 10% of respondents reported being victims of any security risk related to digital payments, which indicates a relatively low incidence of security issues in the realm of digital payment services, but also highlights the need for raising awareness about risks and improving security practices.

#### 4. Conclusion

Research conducted in this work indicates that digital payment services in Serbia provide a relatively high level of security, although full protection has not yet been achieved. The risks in this field are largely recognized but rarely manifest in practice, pointing to the effectiveness of existing security mechanisms. However, the presence of potential risks remains a barrier to the wider adoption of these services, particularly due to concerns about data protection. Furthermore, the existing partial trust in the institutions offering these services highlights the necessity for more sophisticated strategies to enhance security. Accordingly, further efforts in improving security mechanisms, along with educating about risks, can enhance safety and reduce potential threats in the context of digital payment services.

#### References:

- [1] T. M. Bach, W. V. da Silva, A. Mendonca Souza, C. Kudlawicz-Franco, C. P. da Veiga, Online customer behavior: perceptions regarding the types of risks incurred through online purchases, Palgrave Communications, United Kingdom, 2020. doi: 10.1057/s41599-020-0389-4
- [2] A. M. Franciska, Dr. S. Sahayaselvi, An Overview On Digital Payments, 2017, pp. 2101-2111.
- [3] <https://www.forbes.com/advisor/in/banking/what-is-a-digital-payment-and-how-does-it-work/> , accessed 25.09.2024.
- [4] <https://timesofindia.indiatimes.com/gadgets-news/rise-of-the-growing-threat-of-cyber-frauds-due-to-digital-payments/articleshow/92968264.cms> , accessed 29.09.2024.
- [5] P. Wang, M. C. González, R. Menezes, A. L. Barabási, „New generation of mobile phone viruses and corresponding countermeasures, 2010. doi: 10.48550/arXiv.1012.3156
- [6] I. Vayansky, S. Kumar, Phishing – challenges and solutions, Coastal Carolina University, Carolina, 2018, pp. 15-20.
- [7] <https://echelonprotectiveservices.com/what-is-a-physical-attack-in-cyber-security/> , accessed 29.09.2024.
- [8] <https://datadome.co/guides/scraping/scrapper-crawler-bots-how-to-protect-your-website-against-intensive-scraping/> , accessed 03.10.2024.
- [9] T. K. M. Gudimalla, V. Pavithran, S. Kannan, Survey analysis of cloned SIM card, International Conference on Recent Trends in Computing, Communication and Networking Technologies (ICRTCCNT'19), Kings Engineering College, Chennai, Tamilnadu, India, 2019.
- [10] <https://www.geeksforgeeks.org/what-is-sim-card-cloning/> , accessed 03.10.2024.
- [11] T. P. Bhatla, V. Prabhu, A. Dua, Understanding Credit Card Frauds, Cards Business Review, 2003.
- [12] <https://rs.visa.com/about-visa/newsroom/press-releases/prl-28092023.html> , accessed 14.10.2024.



# Neural networks in pellet combustion control - research report for OZONE 55

Stefan Popovic<sup>1</sup>, Dejan Djukic<sup>1</sup>, Sonja Djukic Popovic<sup>2</sup>, Nebojsa Denic<sup>3</sup>, Dejan Viduka<sup>4</sup>

<sup>1</sup> Faculty of information technology, Marsala Tolbuhina 8, Belgrade, Serbia

<sup>2</sup> Faculty of Mathematics, Studentski trg 13, Belgrade, Serbia

<sup>3</sup> Faculty of Science and Mathematic, Lole Ribara 39, Kosovska Mitrovica, Serbia

<sup>4</sup> Faculty of Mathematics and Computer Sciences, Marsala Tolbuhina 8, Belgrade, Serbia

stefan.popovic@alfa.edu.rs; dejan.djukic@alfa.edu.rs; sonjica27@yahoo.com;

nebojsa.denic@pr.ac.rs; dejan.viduka@alfa.edu.rs

## Abstract:

The rapid development of technologies and a faster way of life introduce artificial intelligence and machine learning into all aspects of life. The problems of pollution and global warming have plagued the planet for more than a century, and are the result of excessive consumption of fossil fuels. The last decades have brought innovations in the heating of smaller buildings, heating fuel oil and coal are being shut down, and boilers with automatic gas and biomass heating are being introduced. This significantly reduces pollution, but not enough. Hence the need for greater application of artificial intelligence and machine learning in combustion control in boilers with automatic firing. This paper presents a description of the experimental application of artificial intelligence, machine learning and neural networks to the ATI Terming Ozone 55 boiler and a brief summary of the results obtained.

## Keywords:

neural networks, machine learning, boilers, pellet

## 1. Introduction

Today, when we are faced with uncontrolled pollution, and at the same time the sale of boilers with automatic firing is expanding, we must take care of saving the fuel that the boiler consumes. By increasing the degree of fuel utilization, we reduce the emission of harmful gases into the atmosphere, reduce the consumption of biomass or fossil fuels used, and thus have a positive impact on the environment and increase the quality of living conditions. Neural networks, although they were presented as a mathematical model in the middle of the twentieth century, are experiencing their full expansion only today, at the end of the first quarter of the twenty-first century. The application of neural networks is very broad, there is almost no area in which they did not play a key role in the functioning of automatic control systems, object recognition, decision-making systems and many other areas we are surrounded by.[1] The application of neural networks in the combustion management of boilers with automatic firing is almost negligible. There is almost no research on this topic in papers and practical applications. The application of neural networks itself could be reflected in the automatic calibration of combustion control systems. [2] Therefore, the problem of controlling the operation of boilers with automatic ignition is very current, and the application of neural networks and machine learning to manage the calibration during operation would provide significant savings in fuel consumption, thus contributing to the reduction of harmful effects on the environment and improving the quality of life.

## 2. Problem definition

Safe and stable operation of boilers has been a challenge since ancient times, and combustion is the basis of work, and therefore it is necessary to monitor the state of combustion in order to ensure economical operation of boilers. [3, 4] Increasing economy leads to better fuel utilization, which undeniably results in a positive impact on the human environment. Until now, combustion monitoring systems have determined whether the particles are on or off, but they have not learned from the

measured parameters and improved the combustion parameters. A system that would effectively solve the combustion problem could ensure safe and economical operation of boilers. [2] Although there are systems with better results (GBR model[5]), using knowledge from the field of neural networks for application in combustion control in boilers with automatic ignition would greatly improve the degree of fuel utilization, ensure more stable operation of boilers, which would lead to significant energy savings, and thus would help preserve the environment, reduce the emission of harmful gases, reduce the amount of ash and reduce the consumption of biomass and fossil fuels, which are decreasing every day on planet Earth.

The number of devices connected to different Internet servers around the world is growing rapidly. Gartner's research supports this, according to which there were over twenty billion connected Internet devices in 2015.[6] Today, that number is around 27 billion, while according to research by IHS Market, in 2030 it will be as much as 125 billion. [7] Then, the problem of impaired air quality, which is a consequence of the emission of harmful substances from industrial and thermal power plants, local boiler houses, traffic, which initiates various researches with the aim of analyzing and evaluating the emission of gases and the impact on air quality, and finding solutions for improvement and control. air quality. [8] Global warming and climate changes, impacts on the melting of ice caps, disappearance of plant and animal species - all of the above lead to the inevitability of building cheap hardware and software monitoring solutions that would be an integral part of every energy system.

The above leads to the need to develop a model that will unite artificial intelligence and the basic life needs of the population. The paper before you describes one of the possible applications.

### 3. Research methodology

After researching the literature on the application of neural networks in object recognition and reading the literature, the idea arose that they can also be applied to solving other problems. [1] In parallel with that during the 2019/2020 heating season. a problem was seen with boilers with automatic ignition, which was reflected in unusual shutdown and impossibility of ignition. Then comes the idea that neural networks and machine learning can be applied in troubleshooting. [2]

By monitoring the operation of the boiler during the 2019/2020 heating season, it was determined that whenever pellets from another manufacturer are introduced, there is a stoppage in operation. Then a reason is sought for the uncontrolled shutdown of the boiler and the impossibility of automatic ignition. It is recognized that the automation system that works according to already defined parameters for the quality of wood pellets cannot work with pellets of lower quality. When pre-calibration is carried out in order to burn pellets with higher humidity or lower heat value, the problem of uncontrolled overheating occurs in the event of encountering pellets of better quality. Here comes the idea of introducing a recalibration system during combustion. [9]

Inevitably, there was a need for a system that could take input pellet quality and perform an independent recalibration during operation to eliminate the major stall problem.

In order to adequately approach the finding of causes and possible solutions, the issue of monitoring the operation of the boiler was imposed. The possibility of using one of the existing data recording systems was analyzed. After a comprehensive review of the complexity, it was decided to implement a special purpose system that will collect data from sensors that are embedded in the boiler system in a way that does not disturb the system and has no contact points with combustion. For the experimental data collection, which has been carried out since March 2022 and is still ongoing, an Arduino Uno was used to which K-type temperature probes, humidity meters and gasometers are connected. Temperature probes are placed at the probe points that control the operation of the boiler, but also directly under the firebox, in the smoke channel one meter above the smoke exit from the boiler, as well as for measuring the external and internal air temperature. Humidity sensors monitor the relative humidity of air and pellets. Gas sensors are placed in the chimney to measure the amount and type of gases that are a product of combustion. By May 2022, the system has undergone a series of hardware and software changes to adapt for more accurate and purposeful measurements. Experimental measurements for the research part are carried out from October 2022. [9]

In parallel with the development of the hardware used in the measurements, the selection of a mathematical model that would describe the system began. Herein lies the difficulty that is reflected in

predicting the functioning, order and size of the system. This is important so that they can start collecting data.

By searching for possible mathematical models, we came to the conclusion that a dynamic system with an input and an output where time is discretized can be modeled most easily using the unit difference:

$$y(t) = f(y(t-1), y(t-2), \dots, y(t-D_y), u(t-1), u(t-2), \dots, u(t-D_u)) \quad (1)$$

where knowledge of the model function,  $f$  is the goal of the task of modelling. Here, two important values indicating the size of the problem are  $D_y$ , the order of the autoregressive dependence of  $y(t)$ , and  $D_u$ , the order of the transgressive dependence of  $y(t)$  on  $u(t)$ . It is quite possible that these dependencies stretch back into the infinite past. However, in practice, to have a useful representation of the system, these numbers should be finite. [10]

## 4. Experiment and results

A recurrent neural network was used for data processing. Recurrent neural networks represent a network architecture specialized for processing sequential data, such as natural language sentences and time series. Convolutional networks can also be used for sequential data. The current trend is to reduce the use of recurrent networks and to increasingly insist on the use of convolutional networks even in situations for which recurrent networks seem like a natural solution because they are not easy to train. [11] Nevertheless, in the work, a recurrent neural network was trained in order to check its influence on the operating parameters of boilers with automatic firing.

Note that the recurrent neural network determines one sequence element at a time. We bring a set of measured values to the input and after that we start the calculation. The next mesh is added only after the calculation for the previous mesh has already been performed. A recurrent network maintains a so-called hidden state (vector  $h$ ) and has the current hidden state at each step. For a given input, based on that input and the current hidden state (eg  $h_0$  and  $k_1$ ), it computes a new hidden state as a linear combination of the matrices  $U$  and  $V$  with the vectors of the current hidden state and the current input. An activation function is applied to the obtained result and thus a new hidden state is obtained. An activation function is applied to the obtained result and thus a new hidden state is obtained. After that comes a new set of measurements ( $k_2$ ), based on the current state ( $h_1$ ) and the representation of the set, with the help of the same matrices  $U$  and  $V$ , linear transformations are performed, activation functions are applied and a new state ( $h_2$ ) is obtained, and so on for all measurement sets. At each step, some output can be generated from the hidden state. [9]

## 5. Discussion of the results

The discussion of the results cannot begin before we point out that The neural network training phase incorporates several internal processes and elements, which ensure proper handling and parametrization of the training phase.

1.Number of epochs – Each epoch represents passing the entire dataset forward and backward through the neural network once. Number of epochs is set to 100 for the purposes of the training outlines by this approach.

2.Batch Size – Number of instances processed together before the update of neural network's parameters [12]. Batch size is set to 256, due to the large number of network flow instances.

3.Callback for Learning Rate Reduction – Learning rate influences the size of adjustments made to the model weights based on the calculated gradients. This callback ensures dynamic reduction of learning rate, enhancing and speeding up model's convergence.

4.Callback for Early Stopping – Callback which prevents unnecessary computations and therefore overfitting. After 7 unsuccessful epochs, training phase is terminated and weights from the model's best

performing epoch are selected and restored as they encapsulate model's optimal state achieved in the training phase, for further model's performance evaluation. Complete system recommended and projected by the approach to the methodology given by this paper is presented by Figure 1, encapsulating the entire approach for effective neural network training. [13]

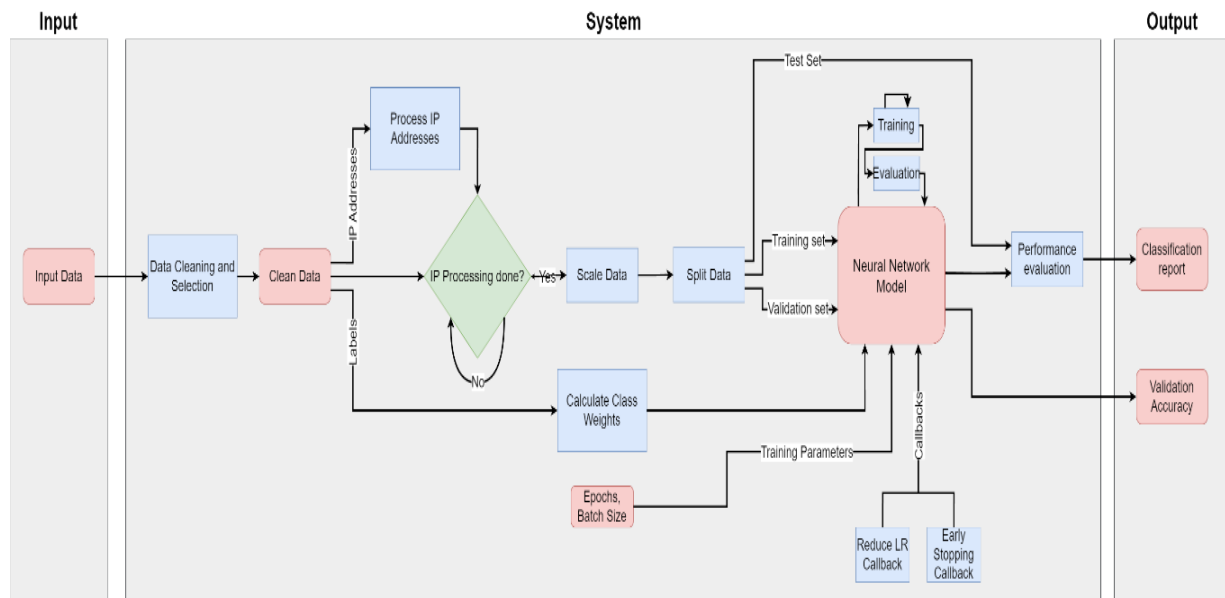


Figure 1: Complete methodology

The carefully designed methodology integrates multiple components to ensure correct and effective neural network's training and evaluation in case of network intrusion detection. The strategic combination of parameters such as the number of epochs, batch size and callbacks ensure both effective and performance-oriented training phase.

By using this model, and after comparing the results, deviations between the computationally obtained hidden states and the actual measured values were seen. By increasing the number of measurement points, the error decreased.

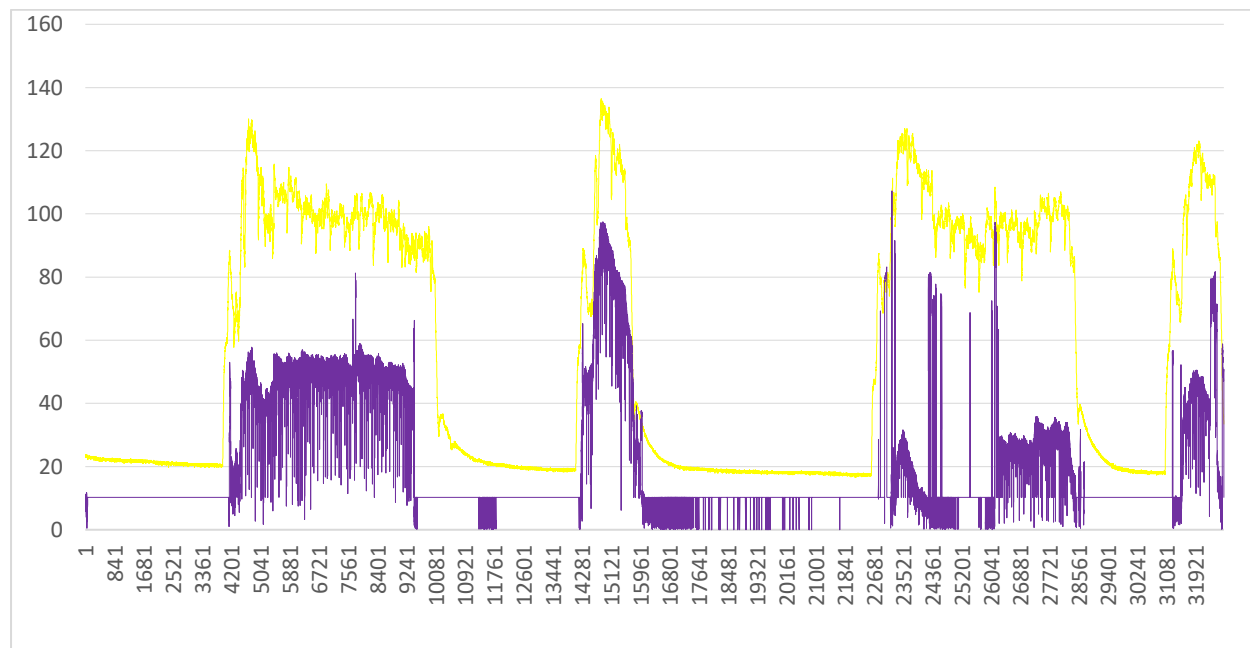


Figure 2: Comparison of measured values and values obtained by calculation

The graphic above shows the relationship between measured values and values obtained by calculation, it can be seen that the calculated value generally follows the measured value, but there is an error. In this research, the neural network obtained an approximate value to the measured value. The error decreased with the number of measurements. And in some cases, the neural network of this type gave results that are not in agreement with the measured results.

## 6. Conclusion

The development of a model of neural networks that will be applied to optimize the control parameters of boilers with automatic firing can lead to a significant reduction in fuel consumption and an increase in the degree of fuel utilization. Forecasting the temperature of the combustion chamber, the energy power of the pellets, the amount of oxygen in the air mixture can significantly improve the operation of the heating system, thereby increasing the comfort of the population and contributing to environmental protection. [9] The importance of the development of algorithms and machine learning today is inevitable. All engineering solutions that are used must contain an artificial intelligence component in their software part, which is also the conclusion of this work.

A big problem in the development of this model is the short period of operation of the boiler, which is contributed by uneven weather conditions. We see the need to develop a model that, in addition to the parameters necessary for fire maintenance and pellet quality monitoring, will also include parameters related to the climatic conditions of the microregion where the measurements are made.

## 7. Proposal for further work and other research

To all researchers working on this topic, we offer, as a team, selfless cooperation and the desire to exchange information and further work on improving the model. In addition, after the completion of the first phase of the project, all obtained results will be posted publicly on the file server.

### Acknowledgment:

Since 2023, this research has been included in the project "Smart cities and modern technologies" registered under the number U03/2023.

### References:

- [1] Stefan Popovic, Lazar Kopanja, Sonja Djukic Popovic and Dejan Djukic (2022). Neural networks and their application in object recognition. Proceedings of 12th International Conference on Applied Internet and Information Technologies (AIIT 2022), Zrenjanin, Serbia.
- [2] Stefan Popovic, Dejan Djukic, Sonja Djukic Popovic and Lazar Kopanja (2022). Preliminary research on the application of neural networks to the combustion control of boilers with automatic firing. Proceedings of 8th Virtual International Conference on Science, Technology and Management in Energy, Belgrade, Serbia
- [3] Fernández-Alemán, J. L., López-González, L., González-Sequeros, O., Jayne, C., López-Jiménez, J. J., Carrillo-de-Gea, J. M., & Toval, A. (2016). An empirical study of neural network-based audience response technology in a human anatomy course for pharmacy students. *Journal of medical systems*, 40(4), 1-12.
- [4] Soltani, A., Battikh, T., Jabri, I., & Lakhroua, N. (2018). A new expert system based on fuzzy logic and image processing algorithms for early glaucoma diagnosis. *Biomedical Signal Processing and Control*, 40, 366-377.
- [5] Wen, H. T., Lu, J. H., & Phuc, M. X. (2021). Applying artificial intelligence to predict the composition of syngas using rice husks: A comparison of artificial neural networks and gradient boosting regression. *Energies*, 14(10), 2932.
- [6] Tomanović Ivana, (2017) Implementation of internet of things in retail development. (English), *InfoM*, Issue 64, p18-25, 8p
- [7] <https://www.spglobal.com/en/research-insights/featured/>

- [8] Borić, J. Goletić, Š.,(2014) Dispersion modelling pollutants emitted from road traffic in the area zenica using software model aermod. (English), Proceedings on Metallic & Nonmetallic Materials, p419-426, 8p
- [9] Stefan Popovic, Dejan Djukic, Sonja Djukic Popovic, Milan Gligorijevic, (2023) Neural networks in pellet combustion control - an overview of the group's research work in 2022/2023, Proceedings of 9th Virtual International Conference on Science, Technology and Management in Energy, Belgrade, Serbia
- [10] D. Djukic (1997), Modelisation des systemes non lineaires, These No 1739, EPFL, Lausanne
- [11] Orvieto, A., Smith, S. L., Gu, A., Fernando, A., Gulcehre, C., Pascanu, R., & De, S. (2023). Resurrecting recurrent neural networks for long sequences. arXiv preprint arXiv:2303.06349.
- [12] Ioffe, Sergey and Szegedy, Christian, (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, arXiv preprint arXiv:1502.0367
- [13] Đorđe Karišić; Milan Čabarkapa, (2024), Neural network-powered intrusion detection system, International Scientific and Professional Conference “ALFATECH “, Belgrade, Serbia

# Revolutionizing education through adaptive technologies

Nemanja Tasić<sup>1</sup>, Dragana Glušac<sup>1</sup>, Vesna Makitan<sup>1</sup>, Miodrag Kovačević<sup>2</sup>, Tamara Milić<sup>1</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Republic of Serbia

<sup>2</sup> University of Novi Sad, Technical College of Applied Sciences, Zrenjanin, Republic of Serbia

nemanja.tasic@tfzr.rs

## Abstract:

The paper aims to systematically present the role that adaptive technologies have in the process of revolutionizing education. The need for change in education is well established and it leads to the process of searching the way of achieving that goal. Searching for solutions in the technology field as well is intuitive but it comes with many benefits and some additional challenges that need to be taken into account. The paper dwells into the complexity of these solutions and tries to gain better insight into the limitations and possibilities that technology offers in education. Results of the research show that major benefits are associated with the personalization of the learning process that helps students acquire knowledge and achieve better academic results.

## Keywords:

*Education, adaptive technologies, change, digitalization, ChatGPT.*

## 1. Introduction

The changes in market demand and the jobs that emerge require that education transforms as well, and adapts to the new needs of the students. One of the ways to achieve that change is through technological solutions and processes that it allows. Learning analytics, artificial intelligence, machine learning, intelligent tutors, adaptive controls, and strong interactive learning materials are examples of emerging technologies that make personalized and adaptable learning possible [1]. The key aspect that needs to be taken into account is to underline the reasons why education needs change and how that change needs to be both aligned and supported with the advancements in the field of technology. The well-established concept for ages is that teaching needs to be adapted to students' interests and personal traits.

Adaptive learning refers to technologies that dynamically adjust to the level or type of course content based on an individual's abilities or skill attainment, in ways that accelerate a learner's performance with both automated and instructor interventions [2].

From the creation of adaptive learning applications to the deployment of adaptive learning systems, the benefits of adaptive learning and the outlook for adaptive learning 3.0's future development persuade that adaptive learning is the best approach to individualized learning and to provide students with a productive, efficient, and enjoyable learning environment [3].

Another way of addressing this topic is perceiving "adaptive learning" as a pedagogical approach that uses cutting-edge technologies, especially machine learning algorithms, to customize teaching methodologies, educational content, and assessment techniques for each student. Based on each learner's performance, preferences, degree of knowledge, and preferred method of learning, it seeks to modify the learning process in real-time [4]. To improve learning experiences in particular educational contexts, educators and practitioners need to be prepared to use AI technology and applications successfully. Furthermore, it is crucial to investigate how conventional abilities like creativity, teamwork, and critical thinking may be incorporated and fostered in AI-driven learning environments [4].

Adaptive learning systems use a range of learning strategies, including item machine learning and artificial intelligence, to personalize the learning process. The Adaptive Learning System was developed to allow students to design their own customized teaching tactics provided they have access to a computer, allowing them to take advantage of a one-on-one teaching model at a cost-effective price and provide each student with their virtual teacher [4].

The increased interest is noted in the paradigm shift in education that shows that it is acknowledged that adaptive learning increases learning efficiency and reduces cognitive overload [5]. It is clear from that point of view that adaptive technologies have a lot to offer in the educational field. The following part of the paper aims to systematically elaborate on how adaptive technologies change the learning process and contribute to a more adequate education system.

## 2. Adaptive technologies in education

### 2.1. Need for change

The traditional educational model has several drawbacks. First of all, the one-size-fits-all strategy ignores the variety of learning preferences and styles that exist within a student body. A standardized curriculum frequently ignores the distinct cognitive styles, strengths, and growth opportunities that each learner possesses. Students may become disengaged and unmotivated as a result of finding particular subjects difficult or boring. Second, the standard methodology doesn't take into consideration how quickly different students pick up concepts. Because of this, the pace of instruction and the various learning rates in a classroom are out of sync, leaving certain pupils unchallenged while others find it difficult to keep up [6]. Instruction has historically taken a "one style fits all" approach in today's typical e-learning environments, meaning that every student is exposed to the same learning processes. Students' varying learning preferences and styles are not taken into consideration in this kind of instruction. Personalized learning, in which education is tailored to each student's unique requirements and learning preferences, has been made possible and facilitated by the advent of e-learning technologies [7].

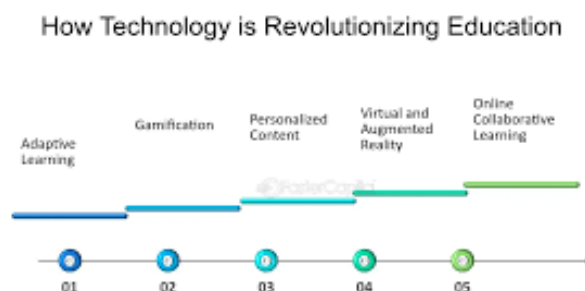


Figure 1. Role of technology in revolutionizing education.

Figure 1 shows how adaptive technologies are the first step for opening other possibilities connected to technological advancements that can further improve and modernize the learning environment which is of great importance for meeting students' needs. It is important to underline why changes should be introduced in the field of education, and also what are some of the most significant barriers.

### 2.2. Benefits

To address the way adaptive technologies can change teaching and learning processes, the benefits should be carefully noted. One of the major benefits is related to the possibility of achieving a personalized learning experience. Algorithms help to determine gaps in student knowledge and to provide recommendations for learning materials and lessons that address those gaps [6]. The findings showed that, in comparison to the non-adaptive system, the students in the adaptive system obtained greater scores in less time [8]. The goal of the personalized learning strategy is to give each student an efficient, effective, and tailored learning path so they may all take part in the process [7]. Adaptive learning programs help students overcome personal learning obstacles by offering courses tailored to their individual needs, including learning preferences [7].



The benefits are present in diverse learning environments and possibility to overcome language barrier by using adaptive learning technologies. It can help enhance students' engagement and learning outcomes. By responding to inquiries, giving information, and making tailored learning recommendations, Amazon Alexa, a voice-activated virtual assistant, provides easily available educational materials and assistance. Learning is becoming more efficient and enjoyable thanks to these AI-driven tools that are changing the educational scene [6]. Students with different learning needs, styles, and capacities can receive specialized support from AI-powered solutions. For instance, AI-powered text-to-speech and speech-to-text programs can help students with impairments by increasing the accessibility of instructional materials. Additionally, AI can promote inclusivity in multicultural classrooms by providing real-time translation and language learning support, thereby removing language barriers [6].

Many studies in specific areas of education noted progress in groups of students in which adaptive technologies were used. Students better understood the content of the class and acquired knowledge better. One of the reasons for these beneficial outcomes is that adaptive technologies increase the intrinsic motivation of students [8].

Active learning, psychological motivation to learn, leveraging past experiences, and effective use of online technology are characteristics of successful and engaged online learners. Additionally, they are skilled at both cooperative and self-directed learning and have outstanding communication skills [7].

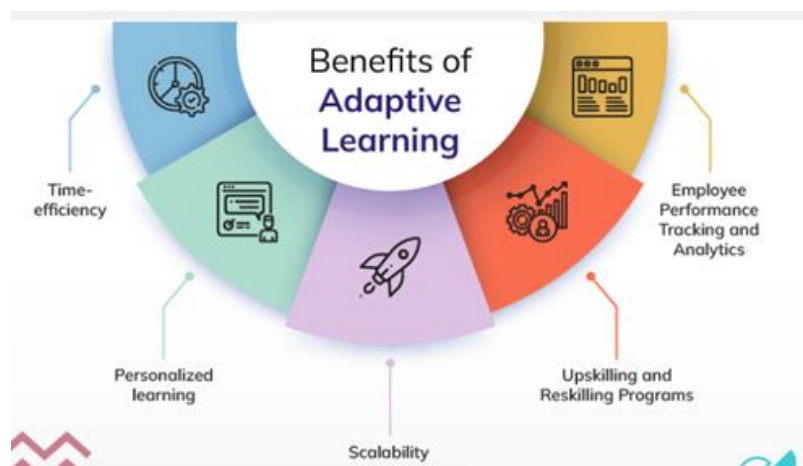


Figure 2. Benefits of Adaptive Learning.

Some of the major benefits are summarized in Figure 2. It also provides insight that learning using adaptive technologies as a student may be significant for a period of employment and further learning endeavors.

Benefits are expected and analyzed also in the scope of the specific types of adaptive technologies. In the new digital era, mobile technology will allow educational institutions to make use of several aspects that will be very helpful to both teachers and students, including flexibility, ubiquity, and portability in learning. Educational institutions and their instructors should create creative teaching strategies to capitalize on students' enthusiasm and the advantages of mobile learning [9]. One of the important benefits associated with the introduction of learning via mobile is that it includes activities that motivate students to actively create new concepts or ideas using what they already know and what they learn. For instance, m-learning enables students to modify current mobile capabilities to suit their requirements, explore their hobbies, and create their learning. Mobile gadgets foster cooperation and communication between students, teachers, and teacher-students as well as provide educational socializing at a distance. Activities that support educators in organizing materials for educational activities. There are programs designed to assist educators in connecting pedagogy and material to the features and objectives of certain mobile applications, and there are mobile resources that could aid educators in comprehending and cultivating new literacies [9].

Recently the focus has frequently on the usage of ChatGPT in classes as a tool to support students' learning and adapt it to their needs. Some of the expected benefits are the following. All things considered, integrating ChatGPT into the educational process gives teachers and students access to

individualized learning opportunities, effective and efficient resource utilization, and adaptive learning strategies that improve learning outcomes for everyone in line with constructivist theory. Students can participate in discovery learning by developing their critical thinking and problem-solving abilities, and teachers can use prompts to speed up the learning process instead of using a didactic method [10]. Promoting individualized and interactive learning, creating prompts for formative assessment exercises that offer continuous feedback to guide instruction, and more are just a few advantages of ChatGPT [11].

### 2.3. Challenges

Aside from the many benefits that are listed, some challenges also need to be addressed so that more real insight into the usage of adaptive technologies in education may be achieved. Bias is one of the main issues since AI systems may inherit biases from biased training data or faulty algorithms, which could lead to unfair treatment and the continuation of current student inequities [6]. Furthermore, privacy is a major worry because AI can gather and analyze enormous volumes of student data, which raises valid questions about how to preserve and use this sensitive data appropriately. AI can provide individualized learning experiences, but it cannot take the place of human connection in the classroom. Concerns around privacy and information security may be the cause of the slow adoption of mobile technology [9].

For children to grow holistically, they need social interaction, emotional support, and feedback from peers and teachers [6]. Some of the limitations of this technology are battery life, storage capacity, and memory limitations. Small screens and low mobile storage capacity restrict the quantity and kind of information that may be shown. Wireless bandwidth is constrained, and as the number of users increases, it may deteriorate [9].

The problems that arise from using this technology are related to the applications' content and the skillfully created m-learning activities. Educational institutions have a significant role to play in the successful implementation of m-learning, even if teachers and students are ultimately responsible for its proper use. The physical features of mobile devices should be taken into account because their tiny and constrained input and output interfaces are undoubtedly a problem [9].

Discussing the mentioned usage of ChatGPT the possibility that data is not accurate is one of the limitations in the usage of this technological advancement in the learning process [10].

All of the challenges, mentioned here and in general, should be understood as the areas that need improvements and that should be understood as the issues that need to be resolved and not the reasons to give up on the usage of adaptive technologies in education.

### 3. Conclusions

Taking into account the emerging need to introduce certain changes in the education process and new technological advancements that make it possible to respond to the needs of students. It is interesting that note that technological changes at the same time represent one of the reasons why the change in education is needed and the solution to how the change may be introduced. Adaptive technologies represent a great way to achieve personalization and an individualized approach to the needs of the students. This is not the only benefit, usage of adaptive technologies also increases the motivation of students, better management of diverse learning environments, and can be understood as a tool that helps in overcoming personal learning barriers. Certain barriers that are also distinguished are present as a part of the process of change and need to be addressed and analyzed in a way to provides solutions and enables technological advancement in education. The contribution of this paper is that systematically presents the main benefits and challenges of introducing adaptive technologies in education. The focus of the paper is not only on the students but teachers' perspectives are also included explaining why some solutions might be hard to incorporate. Additionally, the paper helps in understanding the complexity of the issue and the potential for further research. Future research in this field should focus on new technological solutions and their impact on concrete variables.

**References:**

- [1] D.T. Taylor, M. Yeung, A.Z. Bashet, Personalized and Adaptive Learning, in J. Ryoo, K. Winkelmann (Eds.), *Innovative Learning Environments in STEM Higher Education: Opportunities, Challenges and Looking Forward*, SpringerBriefs in Statistics, 2021, pp. 17-34. [https://doi.org/10.1007/978-3-030-58948-6\\_2](https://doi.org/10.1007/978-3-030-58948-6_2)
- [2] N. Capuano, S. Caballe, Adaptive Learning Technologies, *AI Magazine*, 41 (2), (2020) 96-98. DOI:10.1609/aimag.v41i2.5317
- [3] R.S. Costa, Q. Tan, F. Pivot, X. Zhang, H. Wang, Personalized and adaptive learning: educational practice and technological impact. *SciBrazil*, 14 (3) (2021) <https://doi.org/10.35699/1983-3652.2021.33445>.
- [4] I. Gligorea, M. Cioca, R. Oancea, A.T. Gorski, H. Gorski. R. Tudorache, Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review, *Education Sciences*, 13 (12) (2023) 1216. <https://doi.org/10.3390/educsci13121216>
- [5] S.G., Essa, T. Celik, N.E., Human-Hendricks, Personalized Adaptive Technologies Based on Machine Learning Techniques to identify learning styles: A systematic literature review, *IEEE*, 11 (2023) 392-409.
- [6] B.U. Zaman, Transforming Education through AI benefits, risks, and ethical considerations, *Preprints* (2024) 1-10. <https://doi.org/10.20944/preprints202407.0859.v1>
- [7] H.A. El-Sabagh, Adaptive e-learning environment based on learning styles and its impact on development students' engagement, *International Journal of Educational Technology in Higher Education*, 18 (53) (2021). <https://doi.org/10.1186/s41239-021-00289-4>
- [8] S. Vanbecelaere, K. Van den Berghe, F. Cornillie, D. Sasanguie, B. Reynvoet, F. Depaepe, The effectiveness of adaptive versus non-adaptive learning with digital educational games, *Journal of Computer Assisted Learning* (2019) 1-12. <https://doi.org/10.1111/jcal.12416>
- [9] S. Criollo, A. Guerrero-Arias, A. Jaramillo-Alcazar, S. Lujan-Mora, Mobile Technologies for Education: Benefits and Pending issues. *Applied Sciences*, 11 (9) (2021) 4111.
- [10] T. Rasul, S. Nair, D. Katendra, M. Robin, F.de Oliveira Santini, W.J. Ladeira, M. Sun, I. Day, R.A. Rather, L. Heatcote, The role of ChatGPT in higher education: Benefits, challenges and future research directions, *Journal of Applied Learning & Teaching*, 6 (1) (2023), 1-16.
- [11] D. Baidoo-anu, L. Owusu Ansah, Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning, *Journal of AI*, 7 (1) (2023), 52-62. <https://doi.org/10.61969/jai.1337500>

# The Role of Emerging Technologies in Transforming Supply Chains

Stefan Ugrinov<sup>1</sup>, Edit Terek Stojanović<sup>1</sup>, Dragana Kovač<sup>1</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

*stefan.ugrinov@tfzr.rs; edit.stojanovic@tfzr.rs; dragana.kovac@tfzr.rs*

## Abstract:

Converging emerging technologies are transforming the contours of supply chains by improving efficiency and sustainability, maximising transparency and traceability, reducing costs, and minimising risks. Digitalisation and automation, attained through the Internet of Things (IoT), robotics and artificial intelligence (AI), are creating intelligent networked systems to increase the speed and efficacy of processing data in real time and to automate decisions making by machines and computers both locally and in remote networks. Blockchain technology not only enhances efficiency and traceability in supply chains, but it also increases transparency and trust with secure, irreversible, non-mutating records and ledgers of transactions. AI-driven models are used to optimise demand forecasting, logistics, supplier selection and risk-analysis, thereby ensuring more adaptive and resilient supply chains. Integrating these technologies in intelligent networks has enormous potential to improve supply-chain efficacy and dramatically change the future of global commerce. Governmental investments must be ramped up to support these developments through strategic infrastructural planning, flexible regulatory frameworks and training in educational institutions.

## Keywords:

Digitalization, Blockchain, Artificial Intelligence, Supply Chain, IoT

## 1. Introduction

The integration of emerging technologies into various steps along supply chains is fundamentally transforming the way firms work and engage with global supply chains. In an increasingly competitive environment, digitalisation and automation are becoming key drivers for improving efficiency, reducing costs and for achieving responsiveness to ever-changing market demands. Digitalisation, robotics and automation using artificial intelligence (AI) systems create more intelligent, more responsive, more adaptive and more resilient supply chains. This offers the potential for convenience, improved quality and reliability for consumers, higher value-added for producers and lower costs. Real-time collection of data using the Internet of Things (IoT), combined with monitoring of inventory levels, tracking of shipments and the ability to predict disruptions, help improve the precision and adaptiveness of supply chains [1]. The use of these technologies in managing supply chains minimises human intervention in mundane and repetitive tasks thereby increasing accuracy and reducing inefficiency. The overall consequence is more resilient supply chains.

An emerging application of blockchain is its ability to ensure transparency and traceability across supply chains by creating tamper-proof, sharable records of transactions. In a typical case scenario with a complex, global supply chain that relies on many intermediaries, blockchain technology would capture every step executed along this chain and make it accessible to all of the relevant parties [2]. Through this method, blockchain has the potential to mitigate various risks to interest holders, whether they be the suppliers, manufacturers, wholesalers, retailers, regulators or consumers – such as fraud or counterfeiting, internal or external theft, as well as non-compliance with labour, environmental or other laws. In turn, the fact that every step along the chain can be unequivocally seen would instill greater trust between supply chain partners. Thereby, the lack of any analysis error or perception of collusion would be significantly reduced. The final steps necessary for a consortium to start using blockchains in supply chains include addressing the technical challenges of scaling, interoperability and practical integration with the existing IT systems of participants.

AI and ML, because they can sift through large volumes of historical and current data, are changing how supply chains work. Instead of using demand forecasting summaries culled from data over the last three to six months, these new approaches draw from a wide range and unprecedented volumes of data to feed ML models that can generate more accurate forecasts. AI can help make supply chains more responsive, while reducing waste and inefficiencies [3]. As a result, instead of reacting to sudden, erratic changes in consumer demand and other forces that affect supply, ranging from trade wars to disease outbreaks, AI-assisted demand forecasting helps companies respond in advance to changes. This benefit extends to all other factors of production. AI can efficiently zero in on ideal routes, assign drivers to meet demand and establish acceptable turn times, and help managers pick the most efficient suppliers of raw materials and tools according to different scenarios.

Sustainability is becoming an increasingly key aspect in the supply chain domain, and also one where emerging technologies have the potential to develop greener practices. IoT sensors and AI algorithms optimise the heating/cooling and lighting of warehouses and the entire transportation system – delivered goods and returned goods – to support the humankind's energetic consumption and reduce gross emissions [4]. Blockchain assists companies in monitoring material transparency of suppliers, thus making their materials sourcing more responsible, transparent and sustainable. The use of more sustainable materials, such as biodegradable packaging – a classic example widely used in the UK – is expanding driven by regulatory requirements and customers' expectations. Emerging technologies support circular economy practices, that is, product designs aimed at encouraging reuse, refurbishment or recycling of products, minimising quantities produced and wastage as well as creating new business opportunities.

## **2. The Impact of Digitalization, Automation, and Blockchain on Supply Chain Transformation**

The digitalization and automation of supply chain management have significantly changed business operations, leading to more efficient and responsive processes. As supply chains grow more complex and global, the need for real-time data, precision, and speed has become essential. Digitalization integrates technologies like the Internet of Things (IoT) across the supply chain, transforming traditional methods into interconnected systems. Sensors embedded in products, vehicles, and warehouses provide continuous updates on location, condition, and status, allowing businesses to monitor operations with greater accuracy [5]. This data-driven approach improves decision-making, enabling proactive management of potential disruptions and smoother supply chain operations.

Automation amplifies this transformation by offloading once manual and labour-intensive tasks to machines and AI-enabled systems. Robotics, for instance, play an important role in warehouses, where automated systems manage product picking, packing and sorting. This reduces cycle times, eliminates mistakes and increases efficiency, all while lowering costs. Digitalisation also makes it possible to automate data crunching. Advanced algorithms process vast amounts of information much faster than humans, thereby identifying patterns that may not be immediately visible to the human eye. For example, these insights are crucial in demand forecasting, inventory management and route optimisation benefit from automation [6]. Through digitalisation and automation, today's supply chains are marked by greater efficiency and resilience. Together, these two forces help optimise operations and allow firms to quickly respond to changing demand, supply and external environments.

Blockchain is also crucial in facilitating transparent and traceable supply chains, where there is a need to share information across multiple participants, including manufacturers, suppliers, distributors and retailers. A decentralised, immutable ledger of all supply-chain transactions – with each participant having a single, consistent version of the truth, updated as goods travel through the supply chain – represents a game-changing level of transparency [7]. This is particularly important in industries where the authenticity and origin of goods is crucial – think food and pharmaceuticals. Using blockchain, a product's journey from farm to store could be traced and verified for any consumer.

Blockchain also improves product traceability. If a problem with a product requires companies to recall it, they can use blockchain to trace it all the way back through the supply chain to the source of the problem. This protects consumers and limits the financial and reputational damage to companies.

Blockchain can be paired with other technologies, such as IoT and AI. For instance, IoT devices can record data on the blockchain that updates real-time information about the condition and location of products; combined with AI technology, the data could be analysed to identify problems or inefficiencies, and better decisions made [8]. Of course, there are a few hurdles to widespread blockchain adoption – such as scalability, interoperability and a certain difficulty in collaboration between many stakeholders – many of which companies are still learning how to overcome.

The confluence of digitalisation, automation and blockchain offers companies the potential to build supply chains that are more efficient, transparent, resilient and responsive to customers and stakeholders.

### **3. AI Machine Learning and Sustainability Driving Innovation in Supply Chains**

Artificial intelligence (AI), machine learning (ML) and sustainability are becoming increasingly important to modernising supply chains, particularly in demand forecasting. Demand forecasts are typically based on historical sales using past data as a benchmark for future decisions. But past results will only get you so far in volatile markets. AI and ML use bigger and broader data sets, not only taken from sales, but from websites and social media, to better understand likely future demand. This improves the predictiveness of forecasts and helps brands optimise the amount of inventory they hold, reducing waste and improving the efficiency of their supply chain. Additionally, AI-based models are learning models and over time, they run themselves. The models self-adjust, take into account new data as it becomes available and correct previous predictions in real time. This is especially valuable in supply chains with seasonal demand, or abrupt shifts in consumer tastes [9].

AI and ML also allow for the optimisation of logistics and transportation as it uses variables such as traffic pattern and weather conditions to determine the best possible route for delivery, thereby reducing transportation costs, minimising delays and improving customer satisfaction. AI can also be used to optimise supplier selection, which takes into account various factors such as the reliability of the supplier, cost, environmental impact and so forth. This allows decisions to be taken in the best possible way in accordance with strategic goals. AI is also helping in identifying early signs of supplier distress or potential disruptions through risk management, thereby allowing them to take suitable actions [10]. With AI and ML, the supply chain has become more responsive and resilient as it can anticipate and come up with suitable corrections to handle different circumstances.

Increasingly, concerns around so-called sustainability are front of mind in supply chain management. Indeed, as environmental concerns rise, pressure on businesses to take action increases, with organisations encouraged to reduce carbon emissions and implement more sustainable ways of working. AI, IoT and blockchain have important roles to play here [11]. For example, AI can be used to optimise the use of energy in buildings like warehouses, reducing emissions and keeping operational costs low. IoT devices can be used to monitor the performance of equipment in real-time, improving energy efficiency and identifying issues, such as faults with machinery that need fixed quickly in energy-intensive industries like manufacturing.

Better product and material traceability along the supply chain is a key way by which blockchain technologies can enhance sustainability: it allows companies to trace the responsible and ethical sourcing of raw materials, which is critical for certain industries such as agriculture and fashion. This traceability also helps companies to comply with legislation and expand consumer trust, which is increasingly important as consumers grow ever more demanding of environmentally friendly products. Blockchain supports the transition to a circular economy in various ways [12]. In particular, AI and IoT enable predictive maintenance to extend the product life by optimising upkeep and maintenance. These technologies also help to develop new business models, such as product-as-a-service and remanufacturing, which promote sustainability.

The joint use of AI, blockchain, IoT and other technologies will not only revolutionise the supply chains but also make the supply chains more sustainable. They will harm the environment less, become more effective and more responsive to the growing demands of a sustainable environment [13]. With the increased use of these technologies by businesses, the supply chains will be more resilient. It will open up new avenues for innovative growth for the businesses.

#### 4. Theoretical model

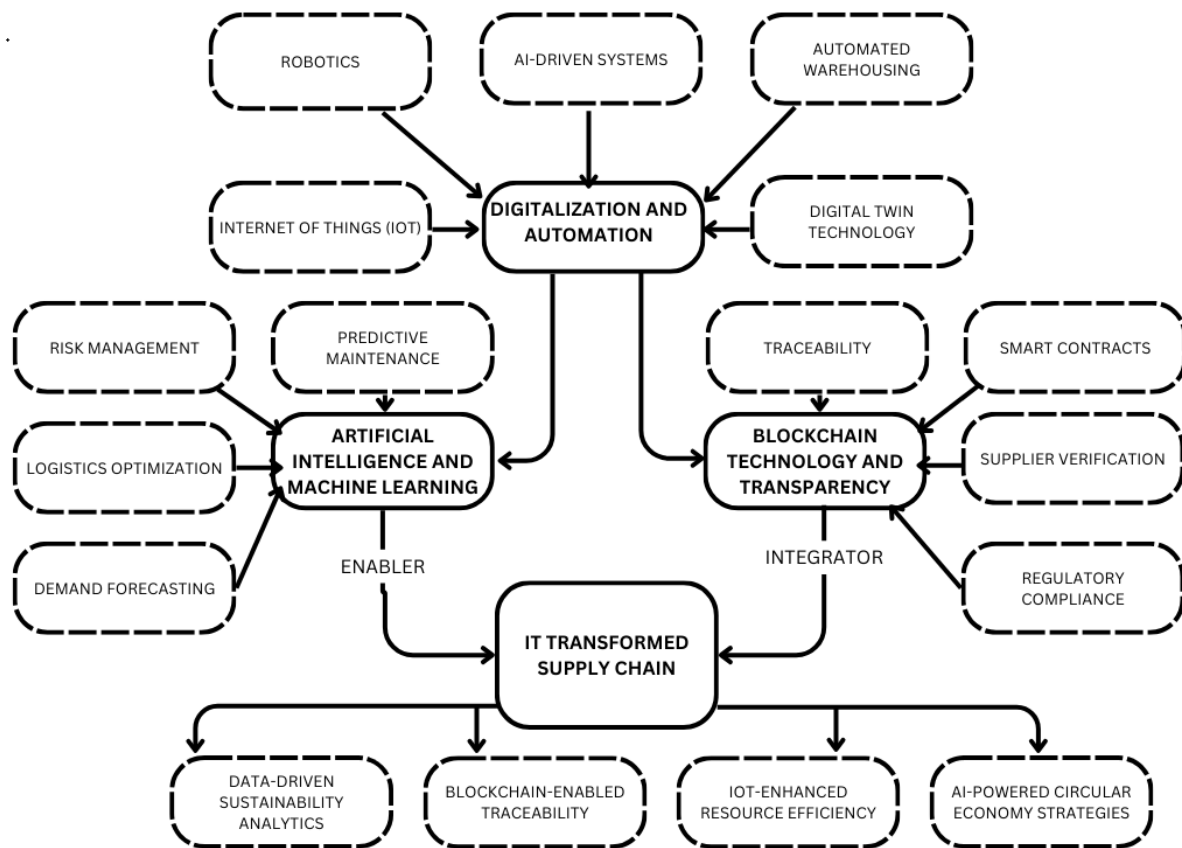


Figure 1 Model for transforming supply chain through emerging technologies

The relationship between digitalisation and automation, blockchain and AI/ML creates an entire ecosystem for supply chain transformation, where these elements and their sub-components complement each other to increase efficiency, transparency, security and flexibility throughout supply chains.

Digitalisation and automation forms the bottom layer of the pyramid, through impactful combination of the Internet of Things (IoT), robotics, artificial intelligence (AI) enabled systems, automated warehousing, and digital twin technology. IoT devices collect data in real time across supply chain processes, which is used for making informed decisions. The data is fed into AI-enabled systems for analysis and optimisation of processes like inventory management and demand forecasting. Robotics automates repetitive and manual tasks for faster and accurate completion of activities such as sorting and packaging. Automated warehousing uses robotics and AI to optimise storage and retrieval systems for increasing the efficiency of material handling and reducing warehousing costs. Digital twin technology enables creation of virtual replicas of physical processes which can be used to simulate and optimise supply chain operations, identify and rectify potential issues before they occur, and implement improved processes without disrupting the actual supply chain workflows.

The sub-elements of blockchain – immutable ledger, traceability, smart contracts, supplier verification, and regulatory compliance – instil a degree of security and transparency. The immutable ledger ensures that the transactions within a blockchain are safely stored and cannot be changed. Traceability allows for the end-to-end tracking of the product, critical in industries where quality and authenticity are not easily identified (such as in pharmaceuticals and food). Smart contracts automate and enforce contractual agreements, whereby specified conditions are executed without human intervention and thereby reduce the potential for dispute. Supplier verification employs the transparent nature of blockchain to authenticate the credentials of suppliers and to ensure compliance with the required standards and ethics. Enhanced regulatory compliance is possible with the accessibility and

tamper-proof records, making audits simpler and ensuring that the required legal standards are adhered to across jurisdictions.

Demand forecasting, logistics optimisation, supplier selection, risk management and predictive maintenance are some of the areas in which AI and ML can significantly enhance the performance of supply chains. For example, AI and ML algorithms can be trained on large, multifaceted datasets to generate accurate demand forecasts that drive optimal levels of inventory. Logistics optimisation can employ AI to calculate the fastest and most efficient routes and schedules for deliveries, taking into account traffic, weather and other factors. The algorithm will then adjust routes and schedules in real time to ensure the most efficient and cost-effective logistics. In supplier selection, AI can assess multiple factors to select the right supplier in real time, depending on the situation, for optimal performance and cost savings. Risk management can benefit from AI's capabilities to identify risks in supply chains. It can continuously monitor and analyse various factors, such as supply chain data, and generate real-time risk alerts, enabling organisations to take pre-emptive actions to minimise the impact of disruptions. Predictive maintenance is a useful tool where ML models can be trained to anticipate component failures before they occur, thus enabling organisations to take action to repair or replace the component before it actually fails. This can save organisations from the financial and operational consequences of unplanned downtime and also extends the lifespan of the assets.

Taken together, these three core components form an extremely dynamic, transparent and resilient supply chain ecosystem. With real-time flow and analysis of data, such a supply chain ecosystem will be better equipped to respond to market changes and disruptions. It will also be in a position to meet the challenges of today's supply chain while further driving innovation and sustainable growth.

## 5. Suggestions and guidelines

The developed model provided insights based on which suggestions and guidelines for emerging technologies in the context of transforming supply chains are highlighted:

- **Governments should invest in digital infrastructure development.** Expanding high-speed internet and supporting IoT networks will enable businesses to fully integrate digitalization and automation in supply chains. This investment will also help rural and underserved areas participate in the global supply chain network.
- **Enterprises should prioritize the integration of AI and ML in demand forecasting and logistics.** Implementing AI-driven systems can significantly improve the accuracy of demand forecasts and optimize logistics, reducing costs and increasing responsiveness. This adoption will help companies remain competitive in a rapidly changing market environment.
- **Enterprises should adopt circular economy practices supported by AI and IoT.** Designing products for longevity and recyclability, and using predictive maintenance, will reduce waste and promote sustainability. This approach will not only align with consumer demand for eco-friendly products but also open up new business opportunities.
- **Collaboration between enterprises and technology providers should be encouraged.** Joint ventures and partnerships can facilitate the development and implementation of customized supply chain solutions, leveraging the latest technologies. This approach will enable businesses to adopt cutting-edge innovations more quickly and effectively.
- **Governments should incentivize sustainable practices in supply chains through tax breaks and grants.** Encouraging the adoption of green technologies and practices will help reduce the environmental impact of supply chains. These incentives will drive businesses to invest in sustainability, contributing to broader environmental goals.
- **Governments and enterprises should collaborate on cybersecurity measures for supply chain technologies.** Protecting data and systems from cyber threats is crucial as supply chains become more digitalized and interconnected. Joint efforts in developing robust cybersecurity protocols will safeguard supply chains from potential disruptions.



## 6. Conclusions

The incorporation of new technologies into supply chains is profoundly transforming how industries operate in terms of efficiency, visibility and sustainability. The increased use of digitalisation and automation has led to highly networked and interconnected supply chains that can process data in real time and make automated decisions. The installation of Internet of Things devices(IoT), robotics, artificial intelligence (AI) systems and automated warehousing have provided supply chains with agility and resilience, allowing them to respond quickly to market fluctuations. These technologies improve efficiency, risk management and reduce costs.

Blockchain technology can be used to create immutable, decentralised transaction records that are impossible to manipulate, creating much needed transparency and traceability while increasing trust among supply chain partners. Smart contracts can also help streamline the processes by automating agreements and eliminating intermediaries.

AI and ML deliver the intelligence that optimises supply-chain operations, including demand forecasting, logistics, supplier selection and risk management. AI-powered predictive maintenance reduces downtime, extend the life of equipment and increase efficiency.

And, given that digitalisation, automation, blockchain and AI are converging to transform supply chains and global commerce in the coming years, these investments in digital infrastructure, regulatory frameworks and skills will be the key to unlocking the potential for these technologies to drive prosperity, sustainability and innovation globally.

### References:

- [1] Andry, Johaness F., Hadiyanto, and Vincensius Gunawan. 2023. "Critical Factors of Supply Chain Based on Structural Equation Modelling for Industry 4.0." *Journal Européen Des Systèmes Automatisés* 56(2):187–94. doi: 10.18280/jesa.560202.
- [2] Aithal, Adithya, Manish Kumar Singh, Papri Ray, and R. Duraipandian. 2021. "An Integrated Approach to Understand Supply Chain Optimization through the Lens of Technology." *Shanlax International Journal of Management* 8(S1-Feb):167–78. doi: 10.34293/management.v8iS1-Feb.3772.
- [3] Kamble, Sachin S., Angappa Gunasekaran, and Shradha A. Gawankar. 2020. "Achieving Sustainable Performance in a Data-Driven Agriculture Supply Chain: A Review for Research and Applications." *International Journal of Production Economics* 219:179–94. doi: 10.1016/j.ijpe.2019.05.022.
- [4] Saberi, Sara, Mahtab Kouhizadeh, Joseph Sarkis, and Lejia Shen. 2019. "Blockchain Technology and Its Relationships to Sustainable Supply Chain Management." *International Journal of Production Research* 57(7):2117–35. doi: 10.1080/00207543.2018.1533261
- [5] Rodchenko, Volodymyr, and Yuliia Prus. 2023. "DIGITAL TECHNOLOGIES IN LOGISTICS AND SUPPLY CHAIN MANAGEMENT." *Facta Universitatis, Series: Economics and Organization* 191. doi: 10.22190/FUEO230517012R.
- [6] Farajpour, Farnoush, Alireza Hassanzadeh, Shaban Elahi, and Mehdi Ghazanfari. 2022. "Digital Supply Chain Blueprint via a Systematic Literature Review." *Technological Forecasting and Social Change* 184:121976. doi: 10.1016/j.techfore.2022.121976.
- [7] Wong, Elton Kee Sheng, Huong Yong Ting, and Abdulwahab Funsho Atanda. 2024. "Enhancing Supply Chain Traceability through Blockchain and IoT Integration: A Comprehensive Review." *Green Intelligent Systems and Applications* 4(1):11–28. doi: 10.53623/gisa.v4i1.355.
- [8] Rejeb, Abderahman, John G. Keogh, and Horst Treiblmaier. 2019. "Leveraging the Internet of Things and Blockchain Technology in Supply Chain Management." *Future Internet* 11(7):161. doi: 10.3390/fi11070161.
- [9] Talwar, Shalini, Puneet Kaur, Samuel Fosso Wamba, and Amandeep Dhir. 2021. "Big Data in Operations and Supply Chain Management: A Systematic Literature Review and Future Research Agenda." *International Journal of Production Research* 59(11):3509–34. doi: 10.1080/00207543.2020.1868599.

- [10] Akbari, Mohammadreza, and Thu Nguyen Anh Do. 2021. "A Systematic Review of Machine Learning in Logistics and Supply Chain Management: Current Trends and Future Directions." *Benchmarking: An International Journal* 28(10):2977–3005. doi: 10.1108/BIJ-10-2020-0514.
- [11] Mandych, Oleksandra, Natalia Babko, Galyna Pasemko, Olga Girzheva, and Oksana Taran. 2023. "Circular economy model: generating benefits for sustainable development." *Economic Analysis* (33(1)):252–57. doi: 10.35774/econa2023.01.252.
- [12] Jagtap, Sandeep, Linh Duong, Hana Trollman, Farah Bader, Guillermo Garcia-Garcia, George Skouteris, Jie Li, Pankaj Pathare, Wayne Martindale, Mark Swainson, and Shahin Rahimifard. 2021. "IoT Technologies in the Food Supply Chain." Pp. 175–211 in *Food Technology Disruptions*. Elsevier.
- [13] Ailane, Amira, Nadia Hamani, Laid Kahloul, and Samir Bourekkache. 2022. "On the Use of Blockchain Technology in Supply Chains: A Brief Review." Pp. 1–8 in *2022 International Symposium on iNnovative Informatics of Biskra (ISNIB)*. Biskra, Algeria: IEEE.

# Regression Gradient Initialization: A New Approach to Weight and Bias Initialization

Dalibor Šeljmeši<sup>1</sup>, Jelena Stojanov<sup>1</sup>, Vladimir Brtka<sup>1</sup> and Dragan Peraković<sup>2</sup>

<sup>1</sup>University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

<sup>2</sup>University of Zagreb, Faculty of Transport and Traffic Sciences, Zagreb, Croatia

*dalibor.seljmesi@tfzr.rs; jelena.stojanov@tfzr.rs; vladimir.brtka@tfzr.rs; dperakovic@fpz.unizg.hr*

## Abstract:

Good weight initialization is important in the convergence and performance of neural networks, particularly in preventing issues such as vanishing and exploding gradients during training. In this paper, we propose a novel weight and bias initialization technique, named Regression Gradient Initialization (RGI), specifically designed for regression tasks. We compare the performance of RGI against widely used initialization methods, including Random, Xavier, and He initialization, across a small regression dataset using a feedforward neural network. Gradient descent and gradient descent with momentum were employed to evaluate the optimization behavior of each method. The results show that RGI produces smoother loss surfaces, leading to more stable and efficient gradient trajectories compared to the other initialization methods. The analysis of gradients and weight evolution further demonstrates the superiority of RGI, particularly in avoiding irregular gradient behavior and enabling more consistent convergence. Our findings suggest that RGI not only ease the challenges of vanishing and exploding gradients but also offers a robust and effective initialization strategy for regression-based neural networks.

## Keywords:

Weight Initialization, Neural Networks, Gradient Descent, Optimization, Loss Surface, Vanishing Gradients, Exploding Gradients

## 1. Introduction

Behind most of today's AI systems are neural networks that solve complex problems in various domains, including, natural language processing, image recognition, and regression analysis. One of the important factors that affects the performance and convergence of neural networks is the initialization of weights and biases. A good initial initialization of a neural network can improve its training speed. The initialization of weights and biases affects the behavior of the gradient descent during training. Also, with a good initialization of the weights, issues such as vanishing and exploding gradients are avoided, which can significantly disrupt the learning process. Exploding gradients occur when weights are initialized with large values, causing gradients to grow exponentially during backpropagation. This can make the training process unstable and the model difficult to optimize. On the other hand, vanishing gradients arise when the weights are initialized too small, causing gradients to diminish as they propagate through the network layers, leading to slow convergence and poor learning. Common initialization techniques include Xavier initialization, He initialization, zero initialization, and random initialization which are discussed in the next section. In this paper, we introduce a novel approach called Regression Gradient Initialization (RGI), specifically designed for regression neural networks. Unlike traditional methods, RGI leverages gradient information from the dataset, initializing weights and biases in a manner that directly addresses the gradient flow challenges inherent in regression tasks. Our technique aims to optimize the initial learning trajectory of neural networks, providing a robust solution to the common issues of vanishing and exploding gradients.

## 2. Previous work

The vanishing and exploding gradient problems have long been a barrier to effective training of deep networks. Several studies have addressed these issues by exploring how different network architectures and initialization strategies impact gradient stability.

Several weight initialization techniques have been developed to address the challenges posed by deep neural networks, particularly the vanishing and exploding gradient problems. In early neural network research, weights were often initialized randomly without much consideration for the depth of the network. This method, while simple, introduced significant challenges, especially as networks grew deeper. Paper [1] highlights how this technique can lead to slow or unstable training, with gradients either diminishing to zero or growing uncontrollably. Although random initialization might work for shallow networks, it often fails in deeper architectures where the flow of information is disrupted by vanishing or exploding gradients.

To ease these problems, Glorot and Bengio proposed the Xavier initialization technique in their paper [2]. Xavier initialization, also known as Glorot initialization, aims to keep the variance of activations constant across layers. Scaling the weights based on the number of incoming and outgoing connections prevents the weights from becoming too large or too small, allowing the gradients to propagate more effectively. This technique showed significant improvements, especially in networks using sigmoid or tanh activation functions, where vanishing gradients were particularly problematic.

Building on the success of Xavier initialization, a new technique called He initialization was introduced in the paper [3]. The initialization, specifically designed for ReLU and other rectified linear activation functions, scales the weights more aggressively by a factor proportional to the square root of 2 over the number of the input units. This prevents neurons from „dying“ during training, a problem that occurs when ReLU units become inactive and stop learning. He initialization has since become a standard for deep networks, particularly those using ReLU or its variants (Leaky ReLU, Parametric ReLU), as it helps maintain the flow of gradients through the network causing exploding values.

In [4], the authors conduct a comprehensive analysis of how certain network designs contribute to gradient issues. They demonstrate that traditional deep feedforward networks are particularly prone to gradient vanishing or explosion, especially in their deeper layers. Their research highlights the effectiveness of architectural innovations such as Residual Networks [5] and Long Short-Term Memory networks [6], which are more robust against gradient-related issues due to their use of shortcut connections and memory mechanisms. These architectures preserve gradient flow, allowing for deeper models that can be trained effectively without suffering from vanishing or exploding gradients.

In [7] authors propose a method that leverages singular value decomposition (SVD) to optimize the parameterization of weights. By ensuring that weight matrices remain well-conditioned, this approach prevents the amplification or suppression of gradients as they propagate through the network. SVD-based initialization methods have been shown to improve convergence, particularly in very deep networks where gradient stability is crucial.

Authors of the paper [8] explore the use of adaptive weight initialization techniques that dynamically adjust during training. By tuning the initialization based on the curvature of the loss surface, these methods can help prevent the vanishing curvature problem, which, like the vanishing gradient problem, leads to slow convergence in deep networks. Adaptive initialization methods offer a more flexible approach that can adjust to the specific characteristics of the data and the network architecture.

As networks become more sophisticated, with architectures like ResNets and WeightNorm [9], specialized initialization techniques have been developed to improve performance. The paper [10] proposes methods specifically designed for these advanced architectures, ensuring that weight norms remain balanced during training.

In most natural applications of machine learning, data arrive incrementally. However, it is common practice to update models sequentially, with each new model starting from the previous one. This practice is generally expected to save time and resources starting from scratch. However, previous studies have shown that although warm-starting achieves similar training losses, it can negatively affect

generalization. The benefits of warm-starting may be bound by complex interactions with hyperparameter tuning and optimization [11].

Paper [12] introduces the concept of hypernetworks, which are meta neural networks capable of generating weights for a main neural network in a fully differentiable process. However, standard weight initialization techniques, such as Glorot and He, when directly applied to hypernetworks fail to yield correctly scaled weights for the main network. Hence, the authors of the above work introduce new initialization strategies for hypernetworks alone, which achieves more stability in the weights of the main network, reduced training loss, and faster convergence.

The authors of paper [13] investigate how weight initialization techniques affect the learning rate, convergence of, and classification accuracy of feedforward neural networks. They have tested Nguyen-Widrow, random, and Xavier initialization techniques coupled with five different activation functions to observe how these combinations can affect the loss rate at the time of training. The research, therefore, compares the performances over a fixed number of epochs and tries to provide insight into the role of initialization in neural networks for the acceleration of training speed and optimization of learning outcomes.

### 3. Proposed Method

We propose a new method for initializing weights and biases in neural networks, particularly targeting regression tasks. This approach, which we call Regression Gradient Initialization (RGI), aims to address issues of gradient instability, including vanishing and exploding gradients, by defining the initial values based on dataset characteristics and the size of the network.

Let  $x \in \mathbb{R}^n$  be a single input sample, where  $n$  is the number of features. In the context of deep learning, the output  $y \in \mathbb{R}^m$  represents the predicted values, where  $m$  is the number of output nodes. The goal is to initialize weights  $w \in \mathbb{R}^k$  and biases  $b \in \mathbb{R}^h$  in such a way that the network starts from a balanced and stable configuration, where  $k$  is the number of weights, and  $h$  number of biases. To achieve that, method directly links the initialization of weights and biases to the dataset, aiming to stabilize gradient flow and improve training performance. The proposed initialization method is based on the equations of the linear regression:

$$y \approx p * x * q, \quad (1)$$

where:

- $y$  is label value from dataset,
- $x$  is one sample from the dataset,
- $p$  and  $q$  are initialization parameters, which can be tuned based on the dataset and network architectures.

Our proposed weight and bias initialization formulas are:

$$w_0 = w_1 = \dots = w_k = \frac{p}{k}, \quad k \in \mathbb{R} \quad (2)$$

$$b_0 = b_1 = \dots = b_h = \frac{q}{h}, \quad h \in \mathbb{R} \quad (3)$$

where:

- $w_0, w_1, \dots, w_k$  are the weights for the network layers, and  $b_0, b_1, \dots, b_h$  are the biases,
- the factors  $p$  and  $q$  ensure that the initial values for both weights and biases are proportional to the number of weights and biases, thus preventing any one feature or neuron from dominating during the initial stages of training.

This method ensures that the gradient updates remain stable during the initial stages of training. By scaling weights as  $\frac{p}{k}$ , we ensure that the magnitude of each weight decreases the number of weights increases, reducing the likelihood of any single feature having too large of an impact on the output. The

bias initialization, scaled as  $\frac{q}{h}$ , similarly balances the output contributions across all neurons, helping maintain a stable flow of gradients across layers.

This initialization ensures that, regardless of the number of neurons or inputs in the network, the activations are balanced, and the gradients during backpropagation are less likely to suffer from vanishing or exploding issues. Regression Gradient Initialization (RGI) scales weights linearly with the number of weights, which is specifically designed for regression tasks and networks were maintaining balanced gradient flow across all layers.

Both  $p$  and  $q$  are derived from the actual dataset, meaning they are data-driven rather than being random or based on abstract statistical principles. This initialization technique aligns the starting values with the structure of the dataset, which can lead to better convergence and training performance, particularly in tasks with continuous outputs like regression.

## 4. Experimental Setup

In this section, we describe the details of the experiment conducted to evaluate the performance of the proposed technique in comparison with standard initialization methods: Random, Xavier, and He initialization. The experiment was carried out using a small regression dataset on a simple feedforward neural network.

### 4.1. Dataset

For the experiment, a small regression dataset was used. The dataset consisted of 42 samples and 1 input features, with continuous target values. Each sample represents an input-output pair, where the network learns to predict the target value based on the input features.

Attributes of dataset:

- Number of samples (N): 42
- Number of input features: 1
- Target variable: Continuous output for regression.

The dataset was split into training and validation sets, with 80% of the data used for training and 20% for validation.

### 4.2. Neural Network Architecture

A simple feedforward neural network with one hidden layer was used. The architecture is described below:

- Input layer: 1 neuron corresponding to the input features.
- Hidden layer: 3 neurons, activated using LeakyReLU.
- Output layer: 1 neuron (for regression) without activation.

The network was designed to be small enough to avoid overfitting, but complex enough to observe differences in initialization performance. Figure 1 represents the neural network:

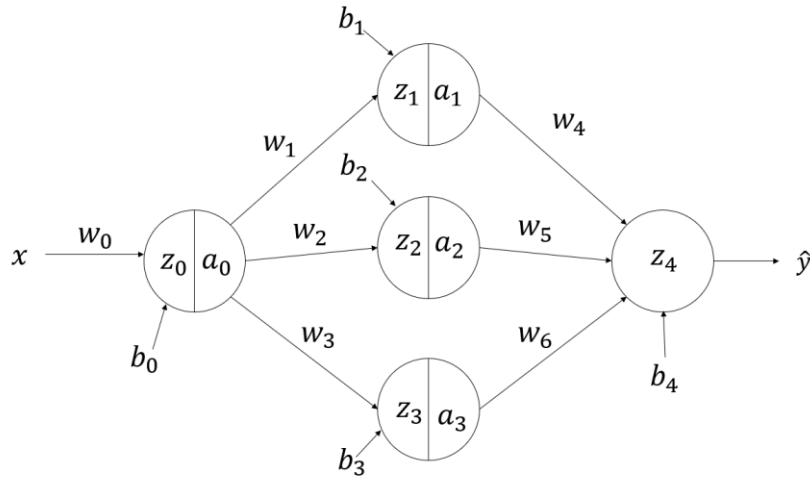


Figure 1: Neural network used for experiment

### 4.3. Weight and Bias Initialization Methods

The network was initialized using four different methods to compare their performance on the task:

- *Random Initialization*: Weights and biases were initialized randomly using a uniform distribution, from 0 to 1.
- *Xavier Initialization*: Weights and biases were initialized following the Xavier uniform initialization method.
- *He Initialization*: Weights and biases were initialized using the He initialization method, suitable for ReLU activations.
- *Proposed RGI Initialization*: Weights and biases were initialized using the proposed regression gradient method.

### 4.4. Training process

The training of the network was performed using two optimization methods:

- *Gradient Descent*: Basic gradient descent was applied with a fixed learning rate of 0.001.
- *Gradient Descent with Momentum*: A momentum term was added to accelerate convergence and escape local minima. The momentum coefficient was set to 0.9 and fixed learning rate of 0.001..

For both methods, the network was trained for 15 epochs, and the performance was monitored at each epoch.

### 4.5. Evaluation Metrics

The models were evaluated using mean squared error (MSE) as the primary performance metric, given that the task is a regression problem. The MSE is defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (8)$$

where  $y_i$  are the true target values and  $\hat{y}_i$  are the predicted values.

## 5. Results and Discussion

Figure 1 shows the loss surfaces generated for each weight initialization method at the beginning of training. The loss surface guides gradient-based optimization, as it determines the ease with which the optimizer can navigate towards a minimum. Each plot visualizes the loss as a function of two weights (denoted  $w_0$  and  $w_1$ ) from the model.

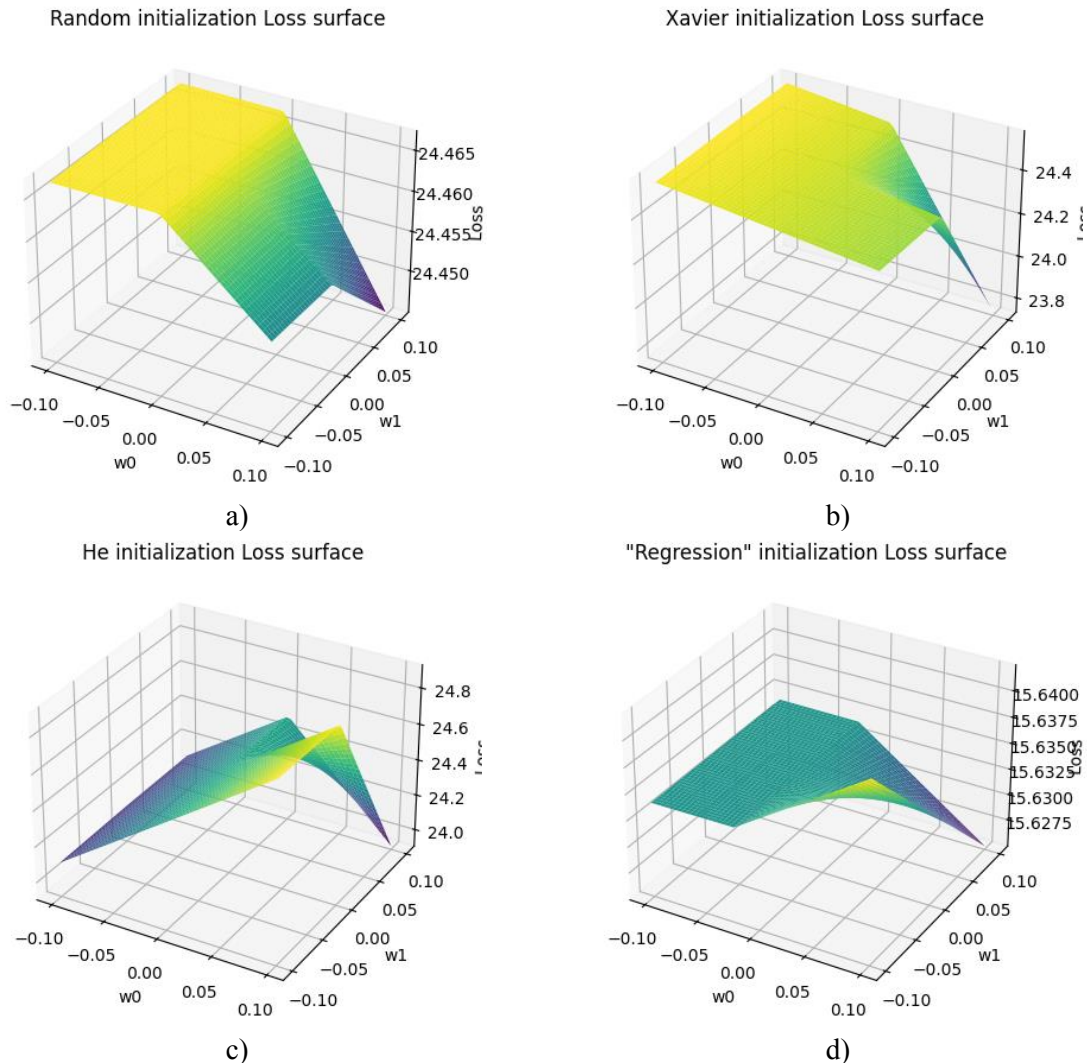


Figure 2: Loss surfaces. (a) Random initialization Loss surface, (b) Xavier initialization Loss surface, (c) He initialization Loss surface, (d) Regression initialization Loss surface (our method)

In Figure 2a) the surface represents the loss landscape generated by Random Initialization. This surface is characterized by an irregular and slightly flat topology, with small gradients and subtle variations across the landscape. The overall loss is relatively high, as indicated by the color map, which ranges around 24.450 to 24.465. Such a surface presents challenges during training, as flat regions with small gradients can slow down the learning process, causing the optimizer to struggle in making meaningful progress. This is typical for random initialization, where weights are selected arbitrarily and often lead to slow convergence or suboptimal solutions.

The loss surface in Figure 2b) corresponds to Xavier Initialization. It exhibits a smoother and more consistent gradient flow than random initialization, with a clear downward slope along one axis. The loss values, which range from approximately 23.8 to 24.4, are lower than those of random initialization,

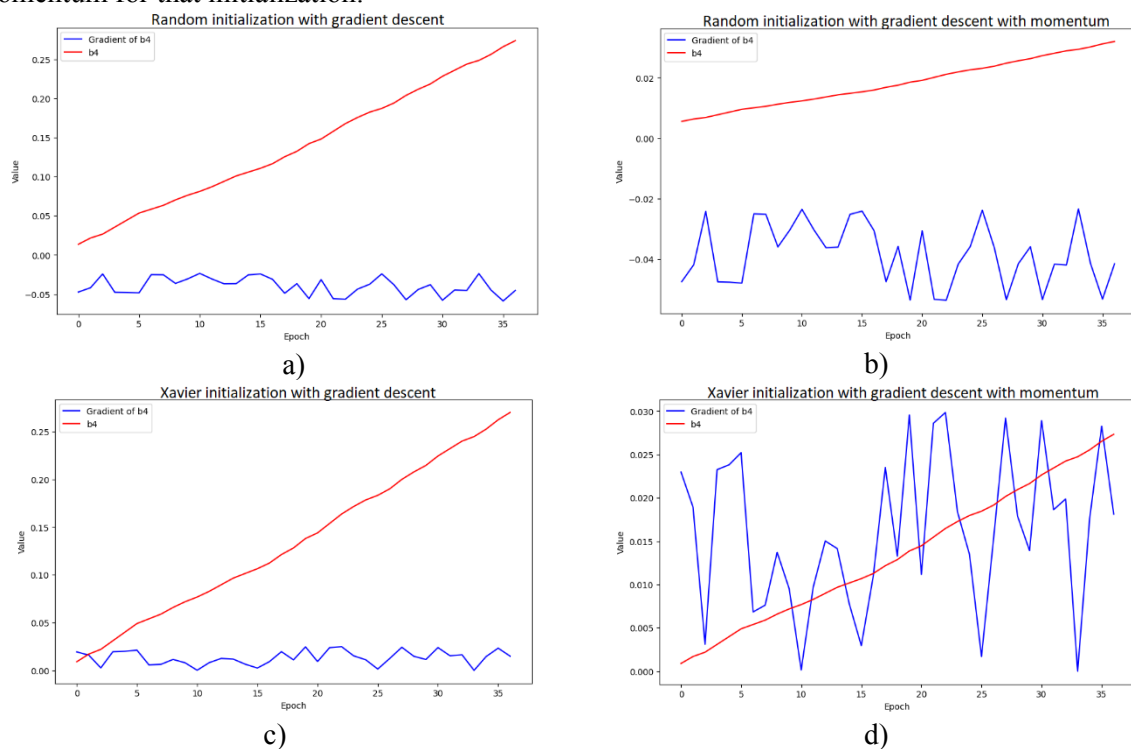


suggesting improved convergence potential. However, there are still regions where the surface is relatively flat, which can lead to the vanishing gradient problem during backpropagation, especially in deeper networks. This method, while better than random initialization, may still struggle with gradient flow, leading to slower training in deeper architectures.

Figure 2c) shows the loss surface for He Initialization, which introduces a more dynamic and sharper landscape. The loss values range from 24.2 to 24.8, indicating a steeper surface compared to both random and Xavier initialization. The sharp transitions and steep slopes in the surface suggest that the gradients in this initialization are more significant, which can accelerate learning in the early stages. However, this method also introduces regions where the gradients can become excessively large, leading to the exploding gradient problem. While He initialization often performs well for networks with ReLU activations, the potential for instability in certain areas of the loss surface can complicate the optimization process, especially when training very deep networks.

Figure 2d) presents the loss surface generated by the Regression Gradient Initialization (RGI) method, the proposed initialization technique. This surface is the smoothest among all four methods, with a loss range from approximately 15.6275 to 15.6400. The relatively lower loss values demonstrate that RGI provides a better starting point for optimization compared to the other methods. The absence of extreme flat or steep regions suggests that the gradients are well-balanced across the surface, reducing the likelihood of both vanishing and exploding gradient issues. This smooth and consistent gradient flow enables the optimizer to traverse the surface efficiently, leading to more stable and faster convergence. The RGI method creates a loss surface that promotes balanced learning dynamics, making it particularly effective for regression tasks.

Figure 3. represents the gradients and bias values of  $b_4$  over epochs for different initialization methods and training algorithms. Each row in the image corresponds to a different weight initialization method, while the two columns represent the performance of gradient descent and gradient descent with momentum for that initialization.



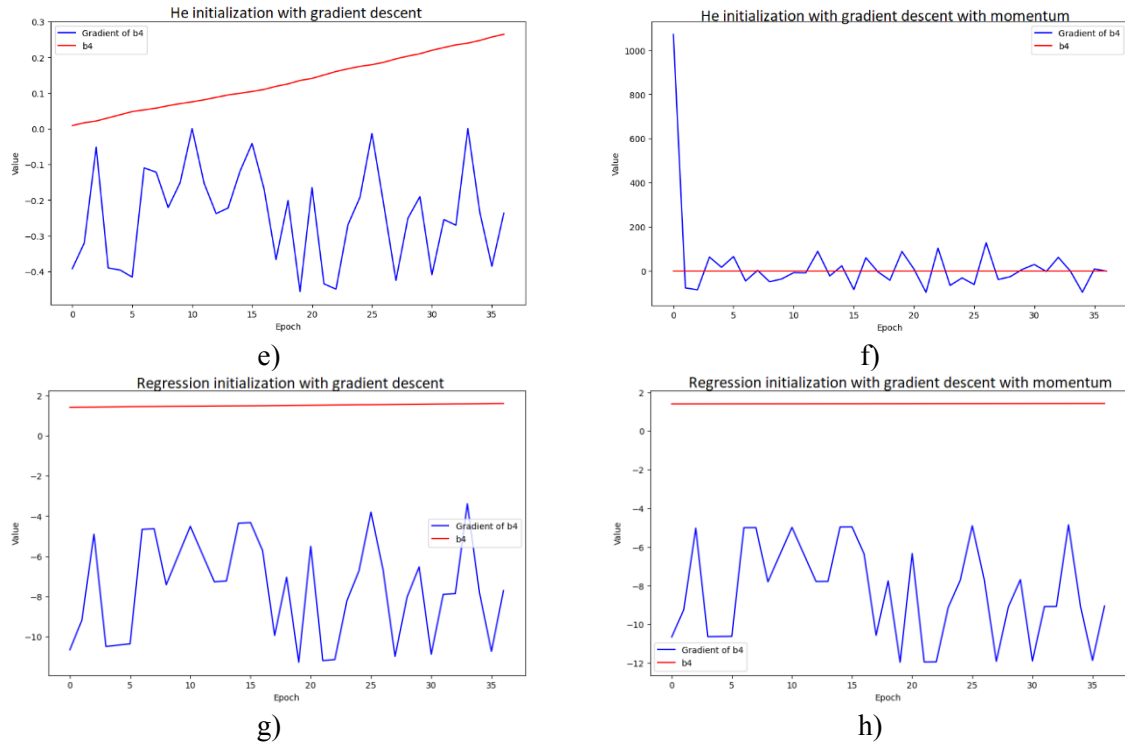


Figure 3: Bias  $b_4$  and gradients. (a) Random initialization GD, (b) Random initialization GDwM, (c) Xavier initialization GD, (d) Xavier initialization GDwM, (e) He initialization GD, (f) He initialization GDwM, (g) Regression initialization GD, (h) Regression initialization GDwM

In Figure 3a) the red line shows the evolution of the bias term  $b_4$  over epochs, while the blue line shows the gradients of  $b_4$ . We observe that the bias gradually increases as the training progresses, indicating steady updates. However, the gradient fluctuates within a small range and remains relatively low, suggesting slow convergence and inefficient learning, possibly due to poor gradient flow.

With the addition of momentum, Figure 3b), the plot shows that the bias term  $b_4$  grows more consistently over time, and the gradients become more stable. However, there are still small shifts, indicating that while momentum helps smooth out the updates, the overall learning process is still not very efficient with random initialization.

In Figure 3c) we see a smoother and more controlled growth of the bias term  $b_4$ . The gradient (blue line) fluctuates within a consistent range and shows less volatility compared to random initialization. The smoother gradient profile suggests that Xavier initialization helps maintain more stable updates, although the gradient magnitudes are still quite small, potentially leading to slower learning in deeper layers.

When momentum is added, Figure 3d), the bias term grows steadily, and the gradient shows more pronounced oscillations compared to the non-momentum case. These oscillations indicate that momentum helps in amplifying the gradients, improving the update dynamics. The larger magnitude of gradient values and smoother convergence in the bias indicate that Xavier initialization with momentum provides more robust learning than simple gradient descent.

In Figure 3e) plot shows larger changes in the gradient of  $b_4$  compared to the previous initializations. The bias term remains almost constant across epochs, with only minor adjustments. This suggests that He initialization causes the gradients to fluctuate significantly, potentially leading to unstable learning. The large gradient magnitudes could be due to the exploding gradient phenomenon, where He initialization can amplify gradients in deeper networks.

When using momentum, Figure 3f), the bias term grows steadily, and the gradient shows extreme shift, particularly at the start of training. The initial spike in the gradient values indicates that the network may experience instability early in training, but the momentum eventually dampens the large oscillations. The addition of momentum somewhat stabilizes the updates, but the large initial gradient values suggest potential issues with this initialization in certain scenarios.

The gradient for  $b_4$  fluctuates within a narrow range, and the bias term remains very stable across epochs, with only minor updates as shown in Figure 3g). This indicates that the proposed Regression Gradient Initialization (RGI) method produces a well-controlled and balanced learning process. The small gradient magnitudes suggest that the network is not encountering vanishing or exploding gradients, allowing for more stable convergence.

With the addition of momentum, Figure 3h), the gradient oscillation remains minimal, and the bias term shows steady, linear growth over time. This indicates that RGI combined with momentum offers the most stable learning dynamics, with neither vanishing nor exploding gradients affecting the learning process. The bias updates are smooth, and the gradient magnitudes are consistently low, making the proposed method highly effective for stable convergence.

The regression-driven approach ensures that the initial values of weights and biases are closer to an optimal solution, allowing the model to converge faster during the training process. This is particularly evident when comparing the gradient descent trajectories and the number of epochs required to reduce the loss function. By starting with more strategically placed initialization values, the RGI method minimizes the time spent navigating flat regions of the loss surface or escaping local minima, as often observed in Random and even Xavier or He initializations.

## 6. Conclusions

In this work, we proposed a novel weight initialization technique, denoted as Regression Gradient Initialization (RGI), and compared its performance against standard initialization methods such as Random, Xavier, and He initialization. Through comprehensive experiments on a regression dataset, we analyzed the behavior of these initialization schemes using gradient descent and gradient descent with momentum.

Our experimental results demonstrate that the choice of initialization significantly impacts the convergence path and the overall behavior of the network during training. The loss surfaces for each initialization revealed distinct landscape characteristics, with the RGI method producing a smoother surface that resulted in more stable and efficient gradient descent trajectories. This indicates that RGI effectively minimizes regions of vanishing and exploding gradients, an issue that is often encountered with traditional initialization methods, particularly in deeper networks.

The analysis of gradient and weight dynamics provided deeper insights into the optimization process. Random and He initialization, while effective in some settings, led to more irregular changes in weight and gradient magnitudes over time, suggesting that the network's parameter space experienced greater instability. Xavier initialization provided a more balanced approach, though it too displayed greater variability in the presence of momentum. RGI, by contrast, yielded consistently smoother gradient trajectories, contributing to a more stable optimization process, especially when momentum was introduced. RGI demonstrated superior stability and convergence properties in our experiments. The smoother gradients and more structured weight evolution suggest that this method can be especially beneficial for regression tasks where a fine-tuned adjustment of weights and biases is important for performance.

Future work could explore how RGI performs on larger and more complex datasets, as well as its potential for adaptation in deep learning architectures, such as convolutional and recurrent networks.

### References:

- [1] D. Sussillo and L. F. Abbott, "Random Walk Initialization for Training Very Deep Feedforward Networks," arXiv preprint arXiv:1412.6558, 2014. Available: <https://arxiv.org/abs/1412.6558>.
- [2] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, vol. 9, 2010. pp. 249-256

- [3] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015. pp. 1026-1034, doi: 10.1109/ICCV.2015.123.
- [4] B. Hanin, "Which Neural Net Architectures Give Rise To Exploding and Vanishing Gradients?" in Neural Information Processing Systems, 2018.
- [5] 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), Las Vegas, NV, USA, 2016. pp. 770-778, doi: 10.1109/CVPR.2016.90.
- [6] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," Neural Computation, vol. 9, pp. 1735-1780, 1997. doi: 10.1162/neco.1997.9.8.1735.
- [7] J. Zhang, Q. Lei, and I. S. Dhillon, "Stabilizing Gradients for Deep Neural Networks via Efficient SVD Parameterization," in Proceedings of the 35th International Conference on Machine Learning (ICML), 2018. pp. 5801-5809.
- [8] A. Orvieto, J. Kohler, D. Pavlo, T. Hofmann, and A. Lucchi, "Vanishing Curvature and the Power of Adaptive Methods in Randomly Initialized Deep Networks," arXiv preprint arXiv:2106.03763, 2021. Available: <https://doi.org/10.48550/arXiv.2106.03763>.
- [9] T. Salimans and D. P. Kingma, "Weight normalization: a simple reparameterization to accelerate training of deep neural networks," in Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16), Curran Associates Inc., Red Hook, NY, USA, 2016. pp. 901-909.
- [10] D. Arpit, V. Campos, and Y. Bengio, "How to Initialize Your Network? Robust Initialization for WeightNorm & ResNets," arXiv preprint arXiv:1906.02341, 2019. Available: <https://doi.org/10.48550/arXiv.1906.02341>.
- [11] Jordan T. Ash and Ryan P. Adams, "On Warm-Starting Neural Network Training," in Advances in Neural Information Processing Systems 33 (NeurIPS), 2020. pp. 3884-3894.
- [12] O. Chang, Lampros Flokas and Hod Lipson, "Principled Weight Initialization for Hypernetworks," arXiv preprint arXiv: 2312.08399, 2023. Available: <https://arxiv.org/pdf/2312.08399>
- [13] Wong, K., Dornberger, R. & Hanne, T. "An analysis of weight initialization methods in connection with different activation functions for feedforward neural networks," Evol. Intel. 17, 2081–2089 2024. Available: <https://doi.org/10.1007/s12065-022-00795-y>

# KNN Algorithm Implementation in Real-World Problem of Water Quality Classification

Kostandina Veljanovska<sup>1</sup>, Aneta Trajkovska<sup>1</sup>, Nikolce Veljanovski<sup>2</sup>

<sup>1</sup> University "St. Kliment Ohridski", Bitola, Faculty of Information and Communication Technologies, Department of Intelligent Systems, Bitola, Republic of N. Macedonia

<sup>2</sup> NIUL "St. Kliment Ohridski", Bitola, Pece Maticevski 39, 7000 Bitola, Republic of N. Macedonia

kostandina.veljanovska@uklo.edu.mk; aneta.trajkovska@uklo.edu.mk; nveljan@yahoo.com

## Abstract:

Technological evolution is increasingly focused on developing intelligent machine learning solutions that improve the efficiency, functionality and adaptability of everyday systems and processes. Effective utilization of algorithms fundamentally enables systems to learn, adapt to environmental conditions and make decisions based on the knowledge, characteristics that define intelligent agents. This capability includes classifying various processes and refining data through user-defined characteristics to produce accurate results. This paper focuses on determining water quality. Water bodies are essential part of our environment and our aim is to contribute to maintain healthy environment for humans, animals and plants. To maintain water quality on high level is very important in order to protect human and all live creatures' health, to avoid the costs related to medical care, costs of productivity loss, and even loss of life. We have implemented the K-Nearest Neighbors (KNN) algorithm to learn from the dataset and develop a classification process based on various parameters. The algorithm quantifies water quality levels based on the concentration of constituent substances. The KNN algorithm is utilized to train the model with the provided data and its performance is evaluated using predefined coefficient values. Results from the implementation of the model yields an accuracy percentage, demonstrating its effectiveness in determining water quality.

## Keywords:

Machine learning, K-Nearest Neighbors algorithm, artificial intelligence, intelligent systems, water quality prediction, water quality classification

## 1. Introduction

Artificial intelligence includes a wide range of technologies, including machine learning, data mining, augmented reality, robotics, logic programming, speech recognition, image processing, ambient intelligence, data visualization and predictive analytics [1]. These methods are applied through various approaches and analyses tailored to specific needs and problems. The focus of machine learning is to automatically learn to recognize complex patterns and make intelligent decisions based on data. The process of learning is shown in Figure 1.

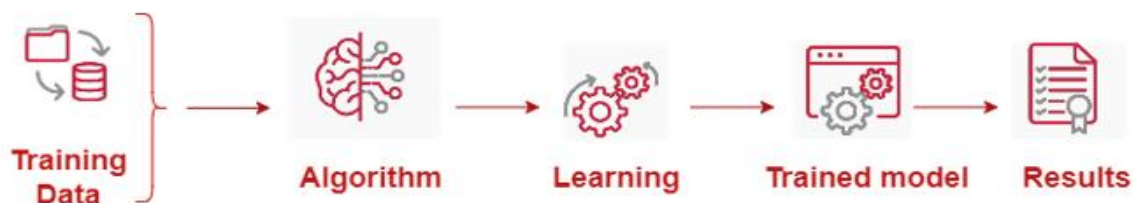


Figure 1: Machine learning progress

Machine learning algorithms can be broadly categorized into three main types: supervised learning, unsupervised learning and reinforcement learning [2]. Each type serves distinct purposes and is applied in various contexts based on the nature of the data and the specific objectives of the analysis [3].

Clean water in every water body is essential for human health, ecosystems, the environment, and the overall wellbeing of our community. Detection and first of all preventing water pollution is essential in order to safeguard water resources, to ensure a sustainable future, and to protect the environment and also endangered population.

In order to protecting world's water resources and environment requires from us to put collective effort. Humans need to foster a sense of responsibility within the community, and as a result help to preserve clean and healthy waterways and aquatic environments for future generations.

Water pollution as a significant worldwide problem can be caused by a variety of sources including industrial waste, inappropriate disposal of various household or institutional chemicals and agricultural runoff. Water bodies which are polluted with dangerous compounds provide significant risks to aquatic ecosystems, human health, and the economy [4].

Within this article, we aim to provide a comprehensive overview of the implementation of KNN algorithm via practical case for determining the water quality based on the concentration of constituent substances in different locations.

The paper's structure incorporates theoretical background of KNN algorithm, the next chapter is discussion about the importance of water quality. In the chapter four, water quality prediction model is presented. Limitations and alternative ways to improve the KNN algorithm in our case are discussed in the next chapter and at the end we conclude with findings during this research.

## 2. KNN Algorithm

There are numerous examples of machine learning algorithms implementation in real-world problems solving and also in water quality determination and prediction [5]. As a powerful tool for solving everyday problems, machine learning was used in this research in order to predict level of water pollution. KNN is a widely utilized supervised learning algorithm known for its simplicity and effectiveness in classification tasks [6]. The core principle is to classify a data point based on the classifications of its nearest neighbors within the feature space. KNN operates on the theory that similar instances tend to be located close to one another in the multidimensional space defined by the input features. When a new data point is introduced, KNN assesses its proximity to the existing labeled instances in the dataset [7]. By evaluating the labels of the K nearest neighbors—where K is a user-defined parameter, the algorithm assigns a classification to the new data point based on the majority label among its neighbors [8]. The algorithm's performance hinges on several factors, including the choice of distance metric (such as Euclidean or Manhattan distance), the selection of K and the feature scaling of the input data [9]. Its interpretability and ease of implementation make it a valuable choice for exploration data analysis and situations where a straightforward classification model is desirable.

The KNN algorithm has demonstrated its utility in a diverse array of real-world applications, ranging from recommendation systems to medical diagnostics. Its ease of implementation, coupled with its capacity to handle complex data relationships, makes KNN a valuable tool in various fields [10]. It is used in healthcare (identifying potential diseases or conditions based on the similarity of new patient data to previously diagnosed cases), agriculture (determining the quality of the soil), finance, social media, detecting the air quality [11], etc. Many popular companies utilize KNN in their systems such as Amazon, Spotify, Google, Netflix, LinkedIn, Uber and others.

## 3. Water Quality

Water pollution appears when different foreign materials contaminate river, lake, sea, ocean or another water body. Often the materials are chemicals, organic, inorganic, biological, radiological, microorganisms or many other substances dissolvable in water which can degrade water quality making the water toxic for people and for environment.

The water, together with the air and the soil are the main elements of our environment. In attempt to address environmental pollution, environmental education is a key factor. [12] Environmental and water pollution education has its main objective to provide learners with knowledge, values and skills that promote the protection and conservation of the environment.

The best thing you can do to prevent water pollution is to educate yourself and your environment about the world's water supply, how you can detect clear water body, understand water pollution and find best means to protect the water resources. Water pollution and water quality education is a very important and effective tool to promote public awareness. We would like to contribute towards this goal by one real world implementation of machine learning algorithm in water quality prediction.

Depending on the aim for water body usage, pollution should be addressed very carefully. Water is a very significant source for all living organisms since all the living organisms need water so it is life itself. Water keeps all of them alive but polluted water is proved to be a very harmful substance. When humans and animals drink polluted water it has serious effects on their health. When their skin is in contact with polluted water, consequences are numerous. Some types of water pollution could be dangerous even for plants. Therefore, it is essential to find models that will react on water pollution quickly and in real-time. Machine learning algorithms as artificial intelligence algorithms implemented in real-time have the capability to gather data form on-site sensors and detect the level of pollution.

## 4. Water Quality Classification Model

This research focuses on classifying water quality based on specific measurement parameters across various locations. The classification framework delineates water into three distinct quality levels: high, medium and low. This assessment incorporates both the geographical context and the concentrations of various substances present in the water, employing weighting factors to evaluate their impact on overall water quality.

### 4.1. Methodology for Analyzing Water Quality Classification

Data collection: the dataset includes measurements of various water quality parameters: nitrate, ammonium, total phosphor, phosphate and chlorophyll. There are other parameters that has to be taken into account, but in order to develop this model and to provide quick results, we determine the scope of our research only to these parameters. Our future research will encompass more parameters and we plan to include parameters joined effect prediction on water pollution taking into account few measurements points next to each another.

Data preprocessing phase: in this phase we have ensured data integrity by data cleaning. It was important to understand well the data and do normalizations to ensure that all variables contributed equally to the classification process. Also adding weighting factors based on the relative importance of determining water quality, based on the following metrics Table 1:

**Table 1:**  
Classes of substances for determining water quality

Parameter	Class High Quality	Medium Quality	Class low Quality	Weighting Factor
Nitrate	0-4,5	4,5-8,5	>8,5	5
Ammonium	0-0,6	0,6-1,6	>1,6	5
TP	0-0,15	0,15-0,7	>0,7	2
Phosphate	0-0,075	0,075-0,35	>0,35	4
Chlorophyll	0-60	60-175	>175	3

Classification framework: water quality was classified into three categories: high, medium, and low, as per the values in Table 1.

Water classification was done based on the Table 1 and taking into account that the parameters that have weighting factor 4 and 5 are crucial. We take water quality value according to those parameters which are prevalent. If one or more of the parameters weighting 4 and 5 have certain value, we take water quality to be of that value.

Algorithm implementation: the KNN algorithm was implemented for the classification process, with utilization of Python programing language, Figure 2.

	monitoringSiteIdentifier	observedPropertyDeterminandLabel	phenomenonTimeSamplingDate	resultObservedValue	ClassQuality
0	EESJA9706000	Ammonium	20140527	0.0700	med
1	IT13PI16	Ammonium	20140527	0.0319	high
2	IT13PE04	Ammonium	20140527	0.0253	high
3	IT13AL13	Ammonium	20140527	0.0183	high
4	IT13GU01	Ammonium	20140527	0.0133	high
5	IT13GU03	Ammonium	20140527	0.0010	high

Figure 2: Analyzing the presence of Ammonium in the water and determining the class quality

Training and Testing Split: dataset was divided into training and testing sets to evaluate the algorithm's performance, demonstrated in Table 2 below.

**Table 2:**

Testing values of the classifier k in the KNN algorithm

Number of data loaded (Database)	Training data	Testing data
23137	18509	4628
1500	1199	300
2500	1999	500
9500	7555	1889

K value determination: various values of k classifier were tested to identify the optimal number of neighbors that yielded the highest classification accuracy, shown in Table 3.

**Table 3:**

Testing values of the classifier k in the KNN algorithm

Parameter	Accuracies	Macro Avg (%)	Weighted Avg (%)
K=1	0.81	0.82	0.82
K=5	0.79	0.79	0.79
K=20	0.74	0.77	0.76

Model evaluation: according to the results obtained, it can be concluded that although we have a change in input data, the KNN algorithm itself has the highest accuracy when the value of the parameter k=1, Figure 3.

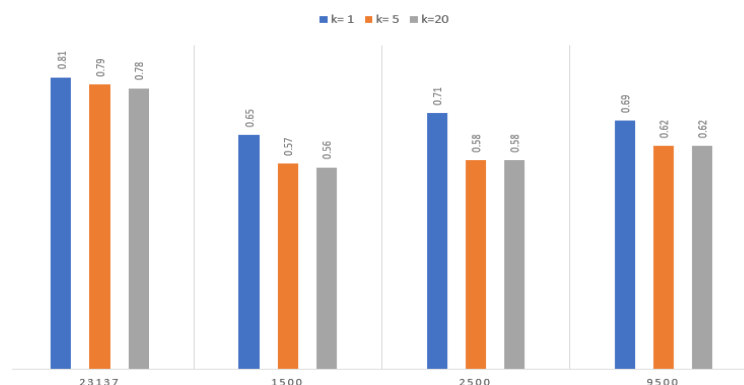


Figure 3: Results obtained from testing the algorithm by loading different numbers of data from the database

## 4.2. Water Classification Based on Date Period and Location

Representation of the results from classification by the algorithm for date 01.07.2020, results shown that the water has medium quality, more details in Table 4.



**Table 4:**Results from the analysis of the values entered on 20200107 for location **SE647907-123283**

<b>Location ID</b>	<b>Nitrate Class</b>	<b>Ammonium Class</b>	<b>Phosphate Class</b>	<b>TP Class</b>	<b>Water Quality (WF)</b>
<b>SE647907-123283</b>	~5.81 med	~1.09 med	~0.54 low	~0.77 low	med

## 5. Limitations and Alternatives to improve the KNN algorithm

Despite its strengths, KNN also has limitations such as computational inefficiency in large datasets, memory intensive and dimensionality issues. The choice of the parameter K is crucial, too small a value can lead to overfitting, while too large a value can result in underfitting. Determining the optimal K often requires cross-validation. As the dataset grows, the K-Nearest Neighbors (KNN) algorithm's can lead to longer processing times and increased resource consumption [13]. There are many alternatives that can be implemented to improve the performance of the KNN algorithm: dimensionality reduction, weighted KNN, using efficient data structures, distance metric variations, hybrid approaches and incremental learning [14].

## 6. Conclusions

Water pollution and water quality education together with scientific effort to determine water quality for each water body is a very important and effective tool to promote public awareness. Our aim is to contribute towards clean, healthy and safety environment by one real world implementation of machine learning algorithm in water quality prediction.

The comprehensive analysis presented in this paper demonstrates the outcomes of the KNN algorithm. KNN algorithm was structured around functions and the implementation of relevant datasets which underwent preprocessing to facilitate the initial learning phase. This step was crucial for enabling the algorithm to perform accurate classifications based on predefined rules embedded in the code. The experimental process revealed how the algorithm's performance was influenced by variations in the data input, demonstrating its adaptability to different datasets. These datasets served as testing resources to refine the classification of water quality into three distinct categories: low, medium, and high. The results showed how the ratio between the weight factors of individual compounds impacted the overall classification, ensuring accurate distribution across the water quality classes. This approach highlights the significance of the influence of each compound on the classification process and its contribution to a more precise evaluation of water quality.

## 7. Acknowledgment

The research work undertaken to study the problem of water pollution and main parameters of water quality, as well as the way the water pollution can be estimated and predicted was performed as a long-time joined effort with our colleagues from Kristianstad University, Department of Education and the Environment. We would like to thank Hristina Bodin, PhD. for their contribution in determining water substances that are important in water quality classification and water quality prediction. They also provided necessary data for the algorithm training. This paper is one small part of the above mentioned research that highlights the power of machine learning in problem of water pollution solution.

### References:

- [1] N. K hl, M. Schemmer, M. Goutier et al. Artificial intelligence and machine learning. Electron Markets 32, 2235–2244 (2022). <https://doi.org/10.1007/s12525-022-00598-0>
- [2] B. Mahesh. "Machine Learning Algorithms - A Review", IJSR (2020), Vol.9, Issue 1, doi:10.21275/ART20203995

- [3] A. Singh, N. Thakur and A. Sharma. "A review of supervised machine learning algorithms", IEEE, 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2016, pp. 1310-1315
- [4] Yadav, Shyma. (2024). Water Pollution: The Problems and Solutions. Science Insights. 44. 1245-1251. 10.15354/si.24.re905.
- [5] Ahmed, A. N. et al., Machine learning methods for better water quality prediction, Journal of Hydrology, Volume 578, 2019, 124084, ISSN 0022-1694.
- [6] A. Moldagulova and R. B. Sulaiman. "Using KNN algorithm for classification of textual documents," 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017, pp. 665-671, doi: 10.1109/ICITECH.2017.8079924.
- [7] M. Steinbach, P.N. Tan. "kNN: k-Nearest Neighbors", (2009), eBook ISBN: 9780429138423
- [8] G. Batista, D. F. Silva. "How k-Nearest Neighbor Parameters Affect its Performance", Argentine symposium on artificial intelligence, doi: b83da505b60e32469152b986cbd6199842403b11
- [9] K. Veljanovska. "Machine Learning Algorithms Comparison", International Journal of Engineering and Technical Research (2017), Vol.7, Issue11, ISSN: 2321-0869 (O) 2454-4698 (P)
- [10] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). <https://doi.org/10.1007/s42979-021-00592-x>
- [11] P. Wang, Y. Zhang and W. Jiang, "Application of K-Nearest Neighbor (KNN) Algorithm for Human Action Recognition," IEEE (2021), pp. 492-496, doi: 10.1109/IMCEC51613.2021.9482165.
- [12] Kılıç, Z., Water Pollution: Causes, Negative Effects and Prevention Methods, Istanbul Sabahattin Zaim University Journal of the Institute of Science and Technology (2021) 3 (1): 129-132.
- [13] Halder, R.K., Uddin, M.N., Uddin, M.A. *et al.* Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *J Big Data* 11, 113 (2024). <https://doi.org/10.1186/s40537-024-00973-y>
- [14] B. Sun, J. Du and T. Gao. "Study on the Improvement of K-Nearest-Neighbor Algorithm," 2009 International Conference on Artificial Intelligence and Computational Intelligence, Shanghai, China, 2009, pp. 390-393, doi: 10.1109/AICI.2009.312.

# Solar-Powered Sensor Station's Energy Consumption Simulator

Dalibor Dobrilović<sup>1</sup>, Jasmina Pekez<sup>1</sup>, Eleonora Desnica<sup>1</sup>, Ivan Palinkaš<sup>1</sup>

<sup>1</sup> University of Novi Sad / Technical Faculty "Mihajlo Pupin" Zrenjanin, Djure Djakovica bb, Zrenjanin, Serbia

*dalibor.dobrilovic@tfzr.rs; eleonora.desnica@tfzr.rs*

## Abstract:

In the era of the rapid growth of IoT, the number of deployed battery-powered wireless sensor nodes rises as well. Therefore the question of building the simulation tool for assessing the energy consumption of solar-powered wireless sensor nodes becomes very important. Considering the rapid development in the field of microcontroller and sensor technologies the flexible simulation tool focused on solar-powered sensor nodes can help in the design and planning of wireless sensor nodes and their deployment.

This paper is focused on an approach to building the Python simulation tool designed to simulate the energy consumption of wireless solar nodes built on open-source hardware. The simulation tool is designed using as a model an existing open-source hardware sensor platform built around ESP32 microcontroller board, and connected sensors. Besides the platform and tool description, the platform results for four scenarios are presented.

## Keywords:

Python simulation tool, solar-powered sensor node, ESP32, sensor node energy consumption

## 1. Introduction

One of the driving powers of the rapid growth of IoT is the variety of roles that IoT devices can have. These devices can be deployed to monitor or manage a variety of systems and in diverse environments. Depending on the IoT type nodes can be deployed in indoor and outdoor environments. In the outdoor environment, these nodes will probably have a problem with the power supply, because not many of them will have access to the power supply in their proximity. If the sensor node is going to be deployed in the open ground, in the fields, or far from the building, it will be probably solar-powered. Therefore, the significance of the proper estimation of energy consumption of sensor nodes rises, especially in the process of their design.

This paper presents the approach in the creation of a Python simulation tool designed to assess the performance of solar-powered sensor stations, as a tool for improving the efficiency of existing prototype platforms. The simulation tool is designed using as a model an existing open-source hardware sensor platform built around ESP32 microcontroller board and connected sensors for monitoring the simple environmental parameters. The simulation is built as an auxiliary tool for improving the efficiency of the same platform, but also to enhance the process of rapid configuration change and introduction of new configurations of wireless sensor nodes.

This paper is structured as follows. After the Introduction, the very brief related work is presented. The platform and simulation overview are presented next. After the platform and tool description, the platform results for four scenarios are presented. Those four scenarios are no sleep-no-solar, sleep-no-solar, sleep-solar-summer, and sleep-solar-winter scenarios. The results are briefly discussed, followed by the conclusion and possible further work definitions.

## 2. Related Work

Numerous studies have explored the application of solar energy, as well as the application and design of simulation tools in various fields of solar energy utilization. Those studies targeted a broad range of solar-powered systems. These studies have been performed over a long period, for the last decade and more. Mazzoni et al. [1] present the methods for developing a simulation tool for solar dish-micro gas

turbine applications which include the solar concentrator, receiver/absorber, micro gas turbine, high-speed generator, and power electronic systems. Makhija al. [2] compared the actual performance of a floating solar plant system in India with simulations from six different commercial tools. The results of their research show that PVsyst and SolarGIS were the most accurate in predicting performance for the canal-top system, while PVsol and SolarGIS–Prospect performed best for the floating system.

Kaleshwarwaret al. [3] examined four system-based (System Advisor Model (SAM), PVsyst, PV\*Sol and Solarius PV) and three online software tools (Helioscope, PV-Watts and PV-GIS) using real-world data from a solar plant installed on a building in Nagpur, India. Buonomano et al. [4] developed a method to maximize the performance of solar-assisted district heating networks. To achieve this goal, a novel dynamic simulation tool for solar thermal field design and optimization has been developed in the Simulink/Simscape environment. In [5] Robles-Enciso et al., presented Java-based simulation tool that allows users to define energy sources and services, test different algorithms, run procedural simulations, and independently optimize energy management. Modeling the solar panels is also very well examined and research in this field dates from a long period ago such as presented in [6].

Although extensive research is performed in the field of solar power system simulation, there is a gap in finding tools for simulating energy consumption in solar-powered wireless sensor nodes. This paper presents the Python-based simulation tool designed to enhance the development of open-source solar-powered sensor device prototypes described in [8] and [9]. This tool is designed to be easy to use and highly flexible to adapt to changes of solar-powered sensor nodes.

### 3. Platform and Simulation Overview

The simulation tool is built using Python and based on open-source prototype sensor nodes described in [8] and [9]. The main components of the platform and simulation tool are presented in Fig. 1 and include:

- Solar Power Subsystem: Covers solar panel ( $81 \times 137$  mm, 1.5 W, 270 mA, 5.5 V), charge controller, Li-Po battery (2.500 mAh), and voltage regulator.
- DHT11 Sensor: This module simulates temperature and humidity readings, incorporating predefined ranges and random variations.
- MCU: is ESP32 microcontroller development board.
- Power Management module: this module implements the power management of the sensor station, and employs power-saving strategies such as sleep modes (Active, Light-sleep, Deep-sleep) and duty cycling.
- Sensor set: Consists of DHT-11 (temperature and humidity sensor), BH1750 light, and UV sensor to generate values derived from solar irradiance data.

The simulation utilizes typical solar irradiance data for typical days, weeks, and months in July and December/January (summer and winter), which can be sourced from reliable references, such as in [7], that reflect average solar irradiance values for these periods. The outcomes of the simulation are displayed through graphs, demonstrating the performance of the solar-powered sensor station across various environmental conditions. MCU ESP32 has 3 connected sensors. Reading of sensor values is triggered each minute, and every 15 min the average data values from all three sensors separately are sent with a WiFi connection.

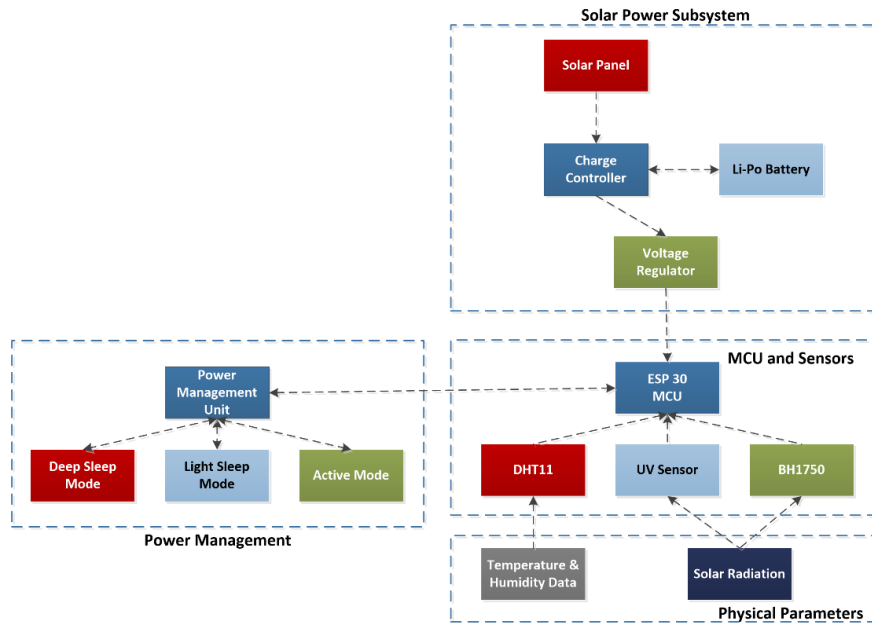


Figure 1: Sensor node and simulation tool components

The main simulation parameters are set as follows:

#### Battery Capacity:

Battery capacity: 2500 mAh

This parameter represents the maximum energy storage capacity of the Li-Po battery declared by the manufacturer.

#### Sleep Mode Currents:

Active\_current: 150 mA  
 Light sleep current: 20 mA  
 Deep\_sleep\_current: 0.01 mA

Active Mode consumes the most energy, as this is the mode in which the device performs sensor measurements and data transmissions. Light Sleep Mode uses reduced power consumption. In this mode the device is idle but can quickly wake when necessary. Deep Sleep Mode is the lowest-power state mode with minimal power usage.

#### Wi-Fi Transmission Parameters:

WiFi transmit current: 300 mA  
 WiFi transmit duration: 2 min  
 WiFi\_transmit\_interval: 15 min

#### Solar Energy Parameters:

Solar radiation summer (SRS): [0, 0, 0, 0, 0, 0, 100, 400, 800, 1000, 800, 600, 500, 600, 800, 1000, 800, 400, 100, 0, 0, 0, 0, 0] W/h  
 Solar radiation winter (SRW): [0, 0, 0, 0, 0, 0, 0, 0, 50, 150, 300, 400, 300, 200, 100, 50, 0, 0, 0, 0, 0, 0, 0, 0] W/h  
 Solar conversion\_efficiency (SCE): 0.2 (20%)  
 Solar\_energy\_day\_summer: SRS \* SCE \* (1.5 / 1000)  
 Solar\_energy\_day\_winter: SRW \* SCE \* (1.5 / 1000)

The summer scenario takes values of radiation during the day in the range between 0 and 1000 Watts/hour, while the winter scenario takes values in the range between 0 and 400 Watts/hour. Sleep

mode impact is designed to simulate the extension of battery life by reducing power consumption during idle periods. Solar conversion efficiency is set to 20 percent.

## 4. Results

The outcomes of the simulation are displayed through graphs. The primary calculations evaluated include:

1. Daily graph with battery level calculation in 24 hours.
2. Weekly graph with battery level calculation in 7 days.
3. Monthly graph with battery level calculation in 30 days.

Four different figures represent the results for three different scenarios. Fig. 3 shows the results for battery level calculations without solar panels and without sleep mode. This simulation scenario estimates the battery operation duration of about 13 hours. Fig. 4 shows the results for battery level calculations without solar panels with implemented sleep modes and in this case battery operation duration of about 2 days. Fig. 4 and Fig. 5 calculate battery consumption for summer (typical temperatures for July), and for winter (typical temperatures for December/January), retrospectively. Those calculus showed that the battery power should be available for continuous sensor node operation.

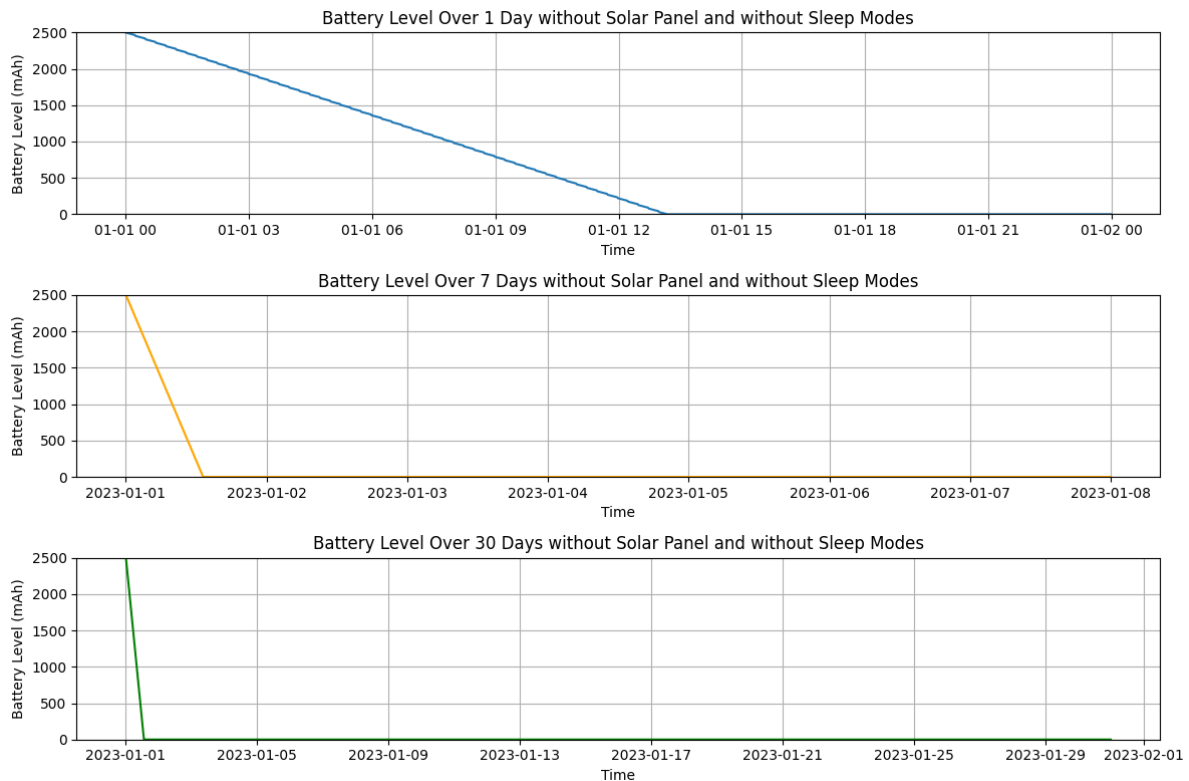


Figure 2: Sensor node simulation for a day, a week, and a month without sleep-mode and without solar panel

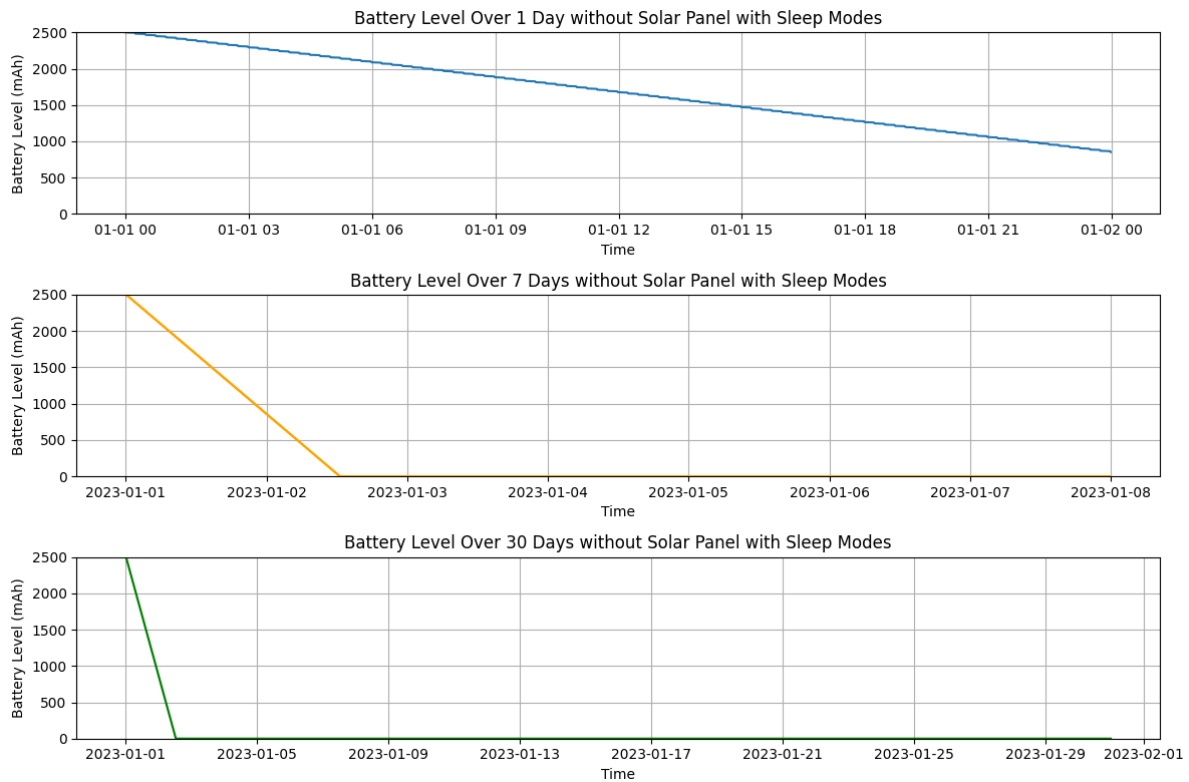


Figure 3: Sensor node simulation for a day, a week, and a month with sleep-mode and without solar panel

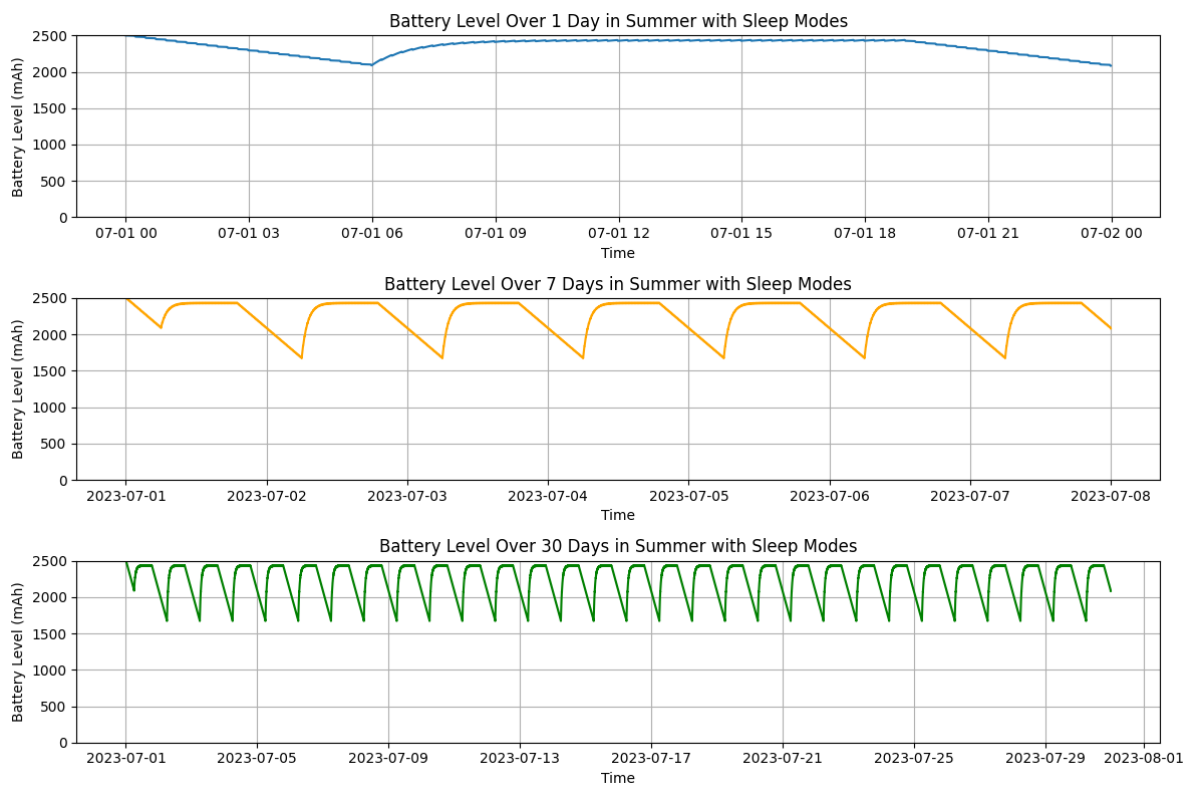


Figure 4: Sensor node simulation for a day, a week, and a month without sleep-mode and without solar panel for July (summer)

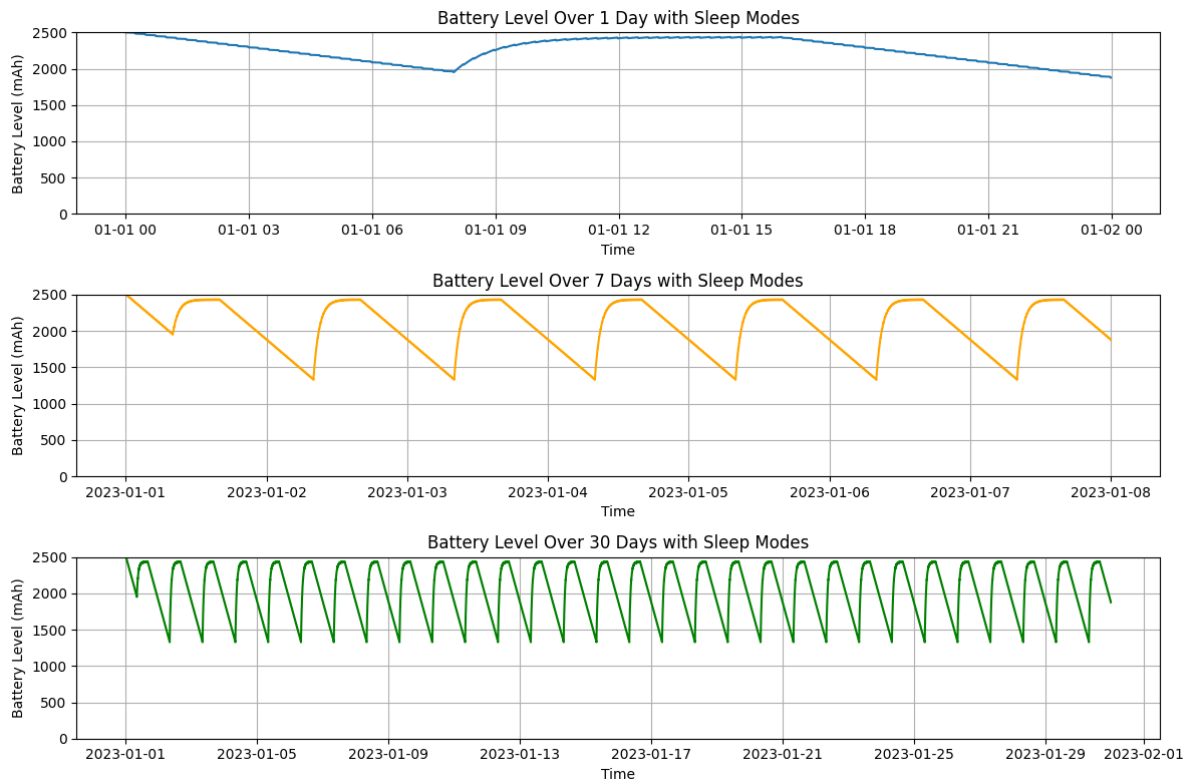


Figure 5: Sensor node simulation for a day, a week, and a month without sleep mode and without solar panel for December/January (winter)

## 5. Conclusion

This paper presents the Python-based simulation tool designed to enhance the development of open-source solar-powered sensor device prototypes described. The simulation tool is applied to four scenarios. Those four scenarios are no sleep-no-solar, sleep-no-solar, sleep-solar-summer, and sleep-solar-winter scenarios. The results are briefly discussed, followed by the conclusion and possible further work definitions.

After the analysis of four scenario results this Python tool proved itself to be efficient in simulating solar energy production and consumption for the specific purpose of powering wireless sensor nodes.

However, the simulation tool can be further improved. It should be compared with the real operating results of prototype sensor nodes that are used as a model for designing this tool. The advanced simulation parameters such as non-linear battery discharge, the influence of battery temperature to battery discharge, the influence of ambient and solar panel temperature on solar output performance, and others should be included in the future.

### Acknowledgment:

This research was conducted through the project “Creating laboratory conditions for research, development, and education in the field of the use of solar resources in the Internet of Things”, at the Technical Faculty “Mihajlo Pupin” Zrenjanin, financed by the Provincial Secretariat for Higher Education and Scientific Research, Republic of Serbia, Autonomous Province of Vojvodina, project number 142-451-3118/2023-01.

### References:

- [1] Stefano Mazzoni, Giovanni Cerri, Leila Chennaoui, A simulation tool for concentrated solar power based on micro gas turbine engines, *Energy Conversion and Management*, Vol. 174, 2018, pp 844-854, <https://doi.org/10.1016/j.enconman.2018.08.059>.



- [2] Amandeep Singh Makhija, Shabbir S. Bohra, Vivek Tiwari, Investigating the performance of water-mounted solar photo-voltaic systems using different simulation tools, *Energy Conversion and Management*, Vol. 322, 2024, 119116, <https://doi.org/10.1016/j.enconman.2024.119116>.
- [3] Akshay Kaleshwarwar, Sarika Bahadure, Validating the credibility of solar simulation tools using a real-world case study, *Energy and Buildings*, Vol. 301, 2023, 113697, <https://doi.org/10.1016/j.enbuild.2023.113697>.
- [4] A. Buonomano, C. Forzano, A. Palombo, G. Russo, Solar-assisted district heating networks: Development and experimental validation of a novel simulation tool for the energy optimization, *Energy Conversion and Management*, Vol. 288, 2023, 117133, <https://doi.org/10.1016/j.enconman.2023.117133>.
- [5] Alberto Robles-Enciso, Ricardo Robles-Enciso, Antonio F. Skarmeta, Sim-PowerCS: An extensible and simplified open-source energy simulator, *SoftwareX*, Vol. 23, 2023, 101467, <https://doi.org/10.1016/j.softx.2023.101467>.
- [6] Villalva, M. G., Gazoli, J. R., & Ruppert Filho, E. (2009). Comprehensive approach to modeling and simulation of photovoltaic arrays. *IEEE Transactions on power electronics*, 24(5), 1198-1208.
- [7] National Renewable Energy Laboratory (NREL), Solar Radiation Data.
- [8] Dobrilovic, D.; Pekez, J.; Ognjenovic, V.; Desnica, E. Analysis of Using Machine Learning Techniques for Estimating Solar Panel Performance in Edge Sensor Devices. *Appl. Sci.* 2024, 14, 1296. <https://doi.org/10.3390/app14031296>
- [9] Dobrilovic, D.; Pekez, J.; Desnica, E.; Radovanovic, L.; Palinkas, I.; Mazalica, M.; Djordjević, L.; Mihajlovic, S. Data Acquisition for Estimating Energy-Efficient Solar-Powered Sensor Node Performance for Usage in Industrial IoT. *Sustainability* 2023, 15, 7440. <https://doi.org/10.3390/su15097440>

# Implementing Master-Detail Interface Design in an MVC .NET Web Application for Gym Membership Management

Nikola Jovanov<sup>1</sup>, Eleonora Brtko<sup>1</sup>, Predrag Novokmet<sup>1</sup>, Milica Mazalica<sup>1</sup> and Aleksandra Stojkov Loncarski<sup>1</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin" Djure Djakovica bb, Zrenjanin, Republic of Serbia

nikola.jovanov@tfzr.rs; eleonora.brtka@tfzr.rs; predrag.novokmet@tfzr.rs; milica.mazalica@tfzr.rs; aleksandra.stojkov@tfzr.rs

## Abstract:

This paper presents the development of an MVC .NET web application designed for gym membership management, with a particular focus on implementing a Master-Detail interface design. The theoretical segment delves into the MVC framework and the crucial role of JavaScript in enhancing user experience through dynamic interactions. It explores the significance of the Master-Detail relationship in efficiently organizing and displaying related data, ultimately improving user productivity. Furthermore, it provides a step-by-step overview of the application development process, detailing the challenges faced and solutions applied along the way. The final product is a user-friendly web application that effectively addresses the needs of gym administrators while showcasing a structured architecture suitable for modern applications.

## Keywords:

Master-Detail Interface, MVC .NET, Web Application, JavaScript Dynamic Rendering

## 1. Introduction

Effective management of gym memberships is crucial for fitness facilities. As competition increases, gyms must ensure they provide efficient services to retain members. A robust system for tracking memberships can streamline administrative tasks and enhance overall operational efficiency for gym staff.

This paper discusses the implementation of an MVC (Model-View-Controller) architecture in a .NET web application designed specifically for administrators. The master-detail interface design is an effective way to display related data in a structured manner. This design allows administrators to view a list of members (master) alongside detailed information for a selected member (detail), facilitating quick access to relevant data. The primary segment of the detail part of this design in the presented application is the list of memberships for the selected member.

JavaScript is employed to dynamically present master-detail relationships within the application, allowing for immediate updates and interactions without the need for page reloads. This approach enhances the user experience for gym staff, making it more intuitive and efficient.

The application features a simple and clear interface, designed for ease of use. While the current CSS styling is rather basic, the structure allows for potential enhancements and improvements in the visual presentation, ensuring that the application can evolve over time to meet user needs.

## 2. Theoretical research

### 2.1. Master-Detail Relationship

A master-detail relationship is a data modeling technique that facilitates the organization and display of related data in a structured manner. In this context, the "master" represents a primary entity, such as a list of members, while the "detail" contains related information specific to the selected master entry. This approach allows users to easily navigate and access relevant data, enhancing the overall efficiency of data management in applications [1].

In relational database systems, master-detail relationships are often implemented using foreign keys. The master table typically contains a primary key, which is referenced by a foreign key in the detail table [2]. This setup ensures referential integrity and allows for efficient querying of related data. By leveraging such relationships, developers can create applications that present detailed information without overwhelming the user with too much data at once [1, 3, 4].

The master-detail interface design enhances user experience by providing quick access to detailed information while maintaining a clear overview of the master data. Users can make informed decisions based on the contextual information provided in the detail view, thereby improving productivity and reducing the time spent searching for specific data points. This design pattern is particularly useful in administrative applications, where users frequently need to access detailed records related to their primary data entries [1, 4].

## 2.2. MVC (Model-View-Controller) Architecture

MVC (Model-View-Controller) is a software architectural pattern widely used in web application development. It separates an application into three interconnected components:

1. Model, which manages the data and business logic;
2. View, which handles the user interface and presentation;
3. Controller, which processes user input and interacts with the model to update the view.

This separation of concerns facilitates easier maintenance, scalability, and testing, making it a popular choice for developers [5].

In the MVC architecture, Views are responsible for rendering the user interface. They receive data from the Controller and display it to the user, allowing for an interactive experience. Views can be dynamically updated based on user input or changes in the underlying data. This component plays a crucial role in ensuring that the application remains user-friendly and visually appealing [5].

Partial Views are a specific type of view in MVC that encapsulates a portion of the user interface. They allow developers to reuse UI components across different views, promoting code reusability and reducing redundancy. By breaking down complex views into smaller, manageable partial views, developers can maintain cleaner and more organized codebases. This modular approach also enhances the maintainability of the application [6].

## 2.3. JavaScript Interactivity

JavaScript is a vital programming language in web development, allowing developers to create interactive and dynamic web applications. It enhances user experience by enabling immediate content updates without requiring a page reload. This feature is particularly valuable in applications where user interaction is frequent, as it provides a seamless experience [7].

One of the core functionalities of JavaScript is event handling, which allows the application to respond dynamically to user actions, such as clicks or key presses. In a master-detail interface, for example, clicking on a member's name triggers an event that fetches and displays detailed information about that member. This not only ensures that the interface remains fluid and user-friendly but also significantly improves usability by facilitating quick access to relevant information. Ultimately, this design pattern enhances productivity for users by providing immediate access to necessary data without the need for page reloads [7, 8].

## 3. Application development

### 3.1. Summary of application functionality

When the application is first accessed, regardless of an attempt to access a specific page on the site, the user is redirected to the login page. There is no registration, as the application is intended for use by supervisors working at the gym's front desk, so login is done using an existing account.

When the administrator logs in, they are redirected to the main page of the application, which serves for displaying member records, as shown in the Figure 1. This page consists of two main

sections – a list of members on the left side and a more detailed description of the selected member, as well as their memberships, on the right side of the screen.

Next to each member in the left table, there is an arrow marked in green for the selected user for display. Above the table, there is a button for creating a new member. In addition to member details, the right side features actions for changing member attributes and deleting them, as well as a button for creating a new membership. Below these actions, there is a table with memberships, where the current active membership is highlighted in green. Additionally, each membership has its own delete button. In the upper right corner, there are buttons for logging out of the account and creating a new user (administrator).

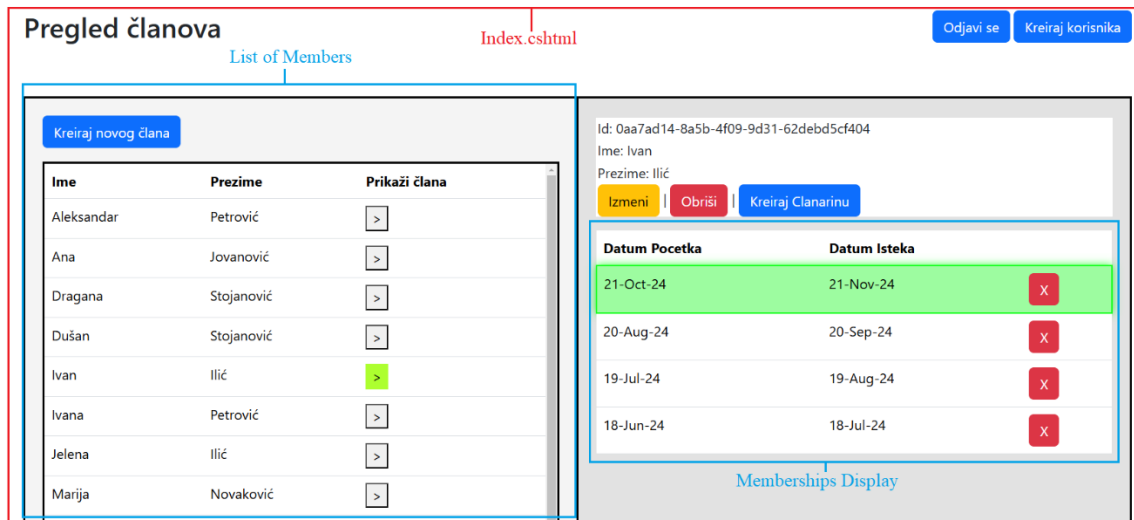


Figure 1: Implementation of master-detail interface, including the names of sections

## 3.2. Implementation of master-detail interface

As previously mentioned, the application consists of a main index page that is divided into two parts:

4. Member List – This section features a compact list of all members, with an arrow next to each member. Clicking on the arrow reveals detailed information about the selected member.
5. Display of Selected Member's Data – This section shows information about the members, actions regarding that entry, as well as their memberships.

This functionality is achieved through the main index view, the partial views for the member list and the memberships of the selected member, and enhanced by JavaScript to manage dynamic interactions and ensure a seamless user experience.

### 3.2.1. MVC Views

From a more practical standpoint, the index page is responsible for displaying all member data. On the left side, it contains a partial view that lists all members, allowing the user to select which member's details to view.

In the right column, the application iterates through each member in the model, displaying their attributes, including ID, first name, and last name. For each member, there are action links for editing, deleting, and creating a membership, providing essential functionality directly from the list.

Immediately following the member details, the partial view for displaying the memberships associated with the selected member is rendered. This structure ensures that users can easily access both member information and their corresponding memberships in a cohesive manner.

Graphic display of mentioned segments is presented in the previously shown Figure 1, while Listing 1 demonstrates code implementation.

```

<div class="kontejner">
 <div class="leva kolona lista-clanova">
 <p>
 @Html.ActionLink("Kreiraj novog člana", "Kreiraj", new { }, new { @class = "btn btn-
primary" })
 </p>
 @Html.Partial("_PrikaziListuClanova", Model.Clanovi)
 </div>
 <div class="desna kolona" id="listaClanovaZaPrikaz">
 @foreach (var item in Model.Clanovi)
 {
 <div id="@item.Id" class="clan-za-prikaz">
 <div style="background-color: white">
 <p class="atribut-clana">Id: @Html.DisplayFor(modelItem => item.Id) </p>
 <p class="atribut-clana">Ime: @Html.DisplayFor(modelItem => item.Ime) </p>
 <p class="atribut-clana">Prezime: @Html.DisplayFor(modelItem => item.Prezime) </p>
 <p class="atribut-clana">
 @Html.ActionLink("Izmeni", "Izmeni", new { id = item.Id }, new { @class = "btn
btn-warning" }) | @Html.ActionLink("Obrisi", "Obrisi", new { id = item.Id }, new { @class = "btn
btn-danger" }) | @Html.ActionLink("Kreiraj Clanarinu", "KreirajClanarinu", "Clanarine", new {
idClana = item.Id }, new { @class = "btn btn-primary" })</p>
 </div>
 @Html.Partial("_PrikaziClanarineZaClana", item.ListaClanarina)
 </div>
 </div>
 }
 </div>
</div>

```

Listing 1: Fragment of Index View

The “\_PrikaziListuClanova” partial view, shown in the Listing 2, is designed to present a scrollable list of members. This view utilizes a table format to clearly display key attributes of each member, including their first name and last name.

Each row in the table contains a button which is linked to the corresponding member's ID. When clicked, this button triggers the display of detailed information about the selected member on the index page. This button has no definition about its actions, since that is a part of the JavaScript code, which will be explained in the later part of the paper.

```

@model IEnumerable<ECT.Web.Models.Clanovi.ClanViewModel>
<div class="lista-za-scollovanje">
 <table class="table" id="listaClanova">
 <tr>
 <th>@Html.DisplayNameFor(model => model.Ime) </th>
 <th> @Html.DisplayNameFor(model => model.Prezime) </th>
 <th>@Html.DisplayName("Prikaži člana")</th>
 </tr>
 @foreach (var item in Model)
 { <tr>
 <td> @Html.DisplayFor(modelItem => item.Ime) </td>
 <td>@Html.DisplayFor(modelItem => item.Prezime) </td>
 <td> <button class="dugme-prikazi-clana" data-id="@item.Id"> > </button></td>
 </tr> }
 </table>
</div>

```

Listing 2: Partial View for list of members

The “\_PrikaziClanarineZaClana” partial view is responsible for displaying a list of memberships associated with a selected member and its implementation can be seen in the Figure 3. This view is structured as a table, which allows for clear and organized presentation of each membership's details. The table includes columns for the start date and expiration date of each membership. For each membership, there is also an action link labeled "X" that allows the administrator to delete the corresponding membership. This integration provides essential functionality for managing memberships directly from the member details view.

```
@model IEnumerable<ECT.Web.Models.Clanarine.ClanarinaViewModel>
<table class="table listaClanarina">
 <tr>
 <th> @Html.DisplayNameFor(model => model.DatumPocetka) </th>
 <th> @Html.DisplayNameFor(model => model.DatumIsteka) </th>
 <th></th>
 </tr>
 @foreach (var item in Model)
 {
 <tr>
 <td id="@item.Id"> @Html.DisplayFor(modelItem => item.DatumPocetka) </td>
 <td> @Html.DisplayFor(modelItem => item.DatumIsteka) </td>
 <td> @Html.ActionLink("X", "ObrisiClanarinu", "Clanarine", new { id = item.Id }, new {
@class = "btn btn-danger" })</td>
 </tr>
 }
</table>
```

Listing 3: Partial View for memberships

### 3.2.2. Enabling dynamic interface management via JavaScript

Upon the DOMContentLoaded event, the script begins execution, ensuring that all HTML elements are fully loaded. It retrieves the elements representing the member display area and the list of members. If either element is missing, an error is logged to the console, and the script halts execution to prevent further errors. This initialization is rather simple and can be seen in Listing 4.

```
document.addEventListener('DOMContentLoaded', () => {
 const listaClanovaZaPrikaz = document.getElementById('listaClanovaZaPrikaz');
 const listaClanova = document.getElementById('listaClanova');
 const clanovi = listaClanovaZaPrikaz.getElementsByClassName('clan-za-prikaz');
 const dugmici = listaClanova.getElementsByClassName('dugme-prikazi-clana');

 if (!listaClanovaZaPrikaz || !listaClanova) {
 console.error('Nedostaju DOM elementi');
 return;
 }
})
```

Listing 4: Initialization of JavaScript code

The next segment, shown in the Listing 5, initializes collections of member elements and buttons. It then iterates through each member element and adds the CSS class “sakriveno” to hide them initially, ensuring that no member details are displayed until a selection is made.

```
for (let clan of clanovi) {
 if (!clan.classList.contains('sakriveno')) {
 clan.classList.add('sakriveno');
 }
}
```

```
}

```

Listing 5: Process of hiding displayed members

An event listener is added to the members list, which listens for click events. If a button for displaying a member is clicked, the ID of that member is retrieved from the button's data-id attribute, and the function “prikaziClana” is called with that ID, triggering the display of the corresponding member's details. Exact implementation can be found in Listing 6.

```
listaClanova.addEventListener('click', (event) => {
 if (event.target.classList.contains('dugme-prikazi-clana')) {
 const id = event.target.getAttribute('data-id');
 prikaziClana(id);
 }
});

```

Listing 6: Event handler for button for displaying a member

The “prikaziClana” function, shown in Listing 7, is responsible for displaying the selected member's details. Initially, it hides all member elements by adding the “sakriveno” class and removing the “prikazano” class. It then identifies the selected member by its ID and removes the “sakriveno” class to show it. Additionally, it updates the visual state of the corresponding button, marking it as active while resetting the state of other buttons to non-active.

```
function prikaziClana(id) {
 for (let clan of clanovi) {
 clan.classList.add('sakriveno');
 clan.classList.remove('prikazano');
 }

 const odabranClan = document.getElementById(id);
 if (odabranClan) {
 odabranClan.classList.remove('sakriveno');
 odabranClan.classList.add('prikazano');

 for (let dugme of dugmici) {
 if (dugme.getAttribute('data-id') === id) {
 oznaciDugme(dugme);
 } else {
 dugme.classList.remove('aktivno-dugme');
 }
 }
 }
}

```

Listing 7: Function for showing a member

The “oznaciDugme” function updates the visual state of the buttons. It first removes the “aktivno-dugme” class from all buttons, ensuring that only one button can be marked active at any time. It then adds the “aktivno-dugme” class to the button that corresponds to the currently selected member, providing a clear visual cue to the user about which member is currently being viewed. This manipulation of CSS classes is displayed in the Listing 8.

```
function oznaciDugme(odabranoDugme) {
 for (let dugme of dugmici) {
 dugme.classList.remove('aktivno-dugme');
 }
}

```

```

 odabranoDugme.classList.add('aktivno-dugme');
}

```

Listing 8: Function for marking a button as selected

The “prikaziPrvogClana” function is designed to automatically display the first member in the list if no member has been selected. It locates the first button for displaying a member and retrieves its ID. If found, it calls prikaziClana with that ID to display the member's details. Listing 9 shows precise implementation of the forementioned function.

```

function prikaziPrvogClana() {
 const prvoDugme = listaClanova.querySelector('.dugme-prikazi-clana');
 if (prvoDugme) {
 const prviId = prvoDugme.getAttribute('data-id');
 prikaziClana(prviId);
 }
}

```

Listing 9: Function for displaying the first member

The “preuzmiProsledjenogClana” function extracts the member ID from the current URL path and is shown in the Listing 10. It splits the pathname into parts, filters out any empty segments, and returns the last segment, which is assumed to be the member ID. This allows the application to load a specific member based on the URL.

```

function preuzmiProsledjenogClana() {
 const path = window.location.pathname;
 const parts = path.split('/').filter(part => part);
 return parts.pop();
}

```

Listing 10: Function for extraction of member ID from URL path

The “isValidGuid” function, shown in the Listing 11, checks if the member ID from the URL conforms to a valid GUID (Globally Unique Identifier) format. This validation ensures that only correctly formatted IDs are processed, preventing errors when displaying member details.

```

function isValidGuid(value) {
 const guidRegex = /^[0-9a-fA-F]{8}-[0-9a-fA-F]{4}-4[0-9a-fA-F]{3}-[89abAB][0-9a-fA-F]{3}-[0-9a-fA-F]{12}$/;
 return guidRegex.test(value);
}

```

Listing 11: GUID validation

If the ID is valid, the application calls the “prikaziClana” function to display the corresponding member's information. If the ID is invalid, the application defaults to showing the first member using “prikaziPrvogClana” presented in the Listing 9, ensuring users always access valid member information. The aforementioned check can be seen in the Listing 12.

```

const id = preuzmiProsledjenogClana();
if (isValidGuid(id)) {
 prikaziClana(id);
} else {
 prikaziPrvogClana();
}

```

Listing 12: Member selection logic



This concludes the member display logic. When the page loads, the application shows the corresponding member based on the provided ID, which is passed if the member was recently created or updated. If no member ID is specified, the application defaults to displaying the first member in the list, ensuring a smooth user experience from the moment the page is accessed.

### 3.2.3. Database Scheme

The database scheme seen in the Figure 2 illustrates the relationship between the "Clanovi" (Member) and "Clanarine" (Membership) tables, forming the foundation for the application's functionality. The "Clanovi" table holds essential information about each member, while the "Clanarine" table contains details about their memberships. It is crucial for these tables to be linked at the lowest layer of the application, as this connection enables the master-detail interface. By establishing a foreign key relationship where the "Clanarine" references the "Clanovi" through the member ID, the application can efficiently display detailed membership information for the selected member.

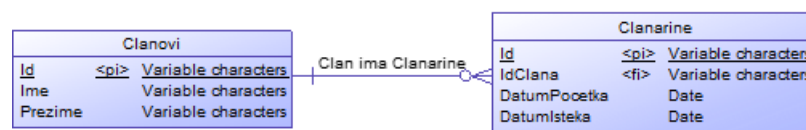


Figure 2: Database tables for members and memberships, including their relationship

## 4. Conclusion and further work

In conclusion, this paper highlights the effective implementation of a Master-Detail interface within a .NET MVC web application for managing gym memberships. The application successfully provides an intuitive user experience by allowing users to view and interact with member data efficiently. The dynamic nature of the interface, facilitated by JavaScript, ensures that users can access relevant information quickly, thereby enhancing productivity.

Looking ahead, there are opportunities for further development, such as the implementation of master-detail entry functionality, which would allow for the input of the first membership directly during the member creation process. Additionally, expanding the application to include member accounts, particularly for mobile devices, would significantly enhance accessibility and usability for gym staff and members alike.

### References:

- [1] Hmel'nov, A., Ruzhnikov, G., Li, T., & Xu, H., "An approach to information export from database for complex master-detail table hierarchies into a single flat table form", CEUR Workshop Proceedings, 2021.
- [2] Silberschatz, A., Korth, H., & Sudarshan, S., "Database System Concepts (Sixth Edition)", McGraw-Hill, 2011.
- [3] Naik, P. G., & Naik, G. R., "JSP Custom Tag For Displaying Master-Detail Relationship in a Hierarchical Grid Control – A Case Study", International Journal of Engineering Applied Science and Technology, 2016.
- [4] Sells, C., & Griffiths, I., "Programming WPF (Second Edition)", O'Reilly Media, 2007.
- [5] Galloway, J., Haack, P., Wilson, B., & Allen, K. S., "Professional ASP.NET MVC 4", Wrox, 2012.
- [6] Ozkaya, M., Fidandan, I., "MVCLang: A Software Modeling Language for the Model-View-Controller Design Pattern", International Conference on Software Technologies, 2020.
- [7] Flanagan, D., "JavaScript: The Definitive Guide (Seventh Edition)", O'Reilly Media, 2020.
- [8] Crockford, D., "JavaScript: The Good Parts", O'Reilly Media, 2008.

# Adaptive Biometric Systems: Challenges and Opportunities

Anita Petreska<sup>1</sup>, Blagoj Ristevski<sup>1</sup>, Ivica Dimitrovski<sup>2</sup>, Saso Nikolovski<sup>1</sup>, Daniela Slavkovska<sup>3</sup>

<sup>1</sup> Faculty of Information and Communication Technologies-Bitola, University "St. Kliment Ohridski"-Bitola, Republic of Macedonia,

<sup>2</sup> Faculty of Computer Science and Engineering, University of Ss. Cyril and Methodius in Skopje, Republic of Macedonia

<sup>3</sup> Children's Hospital Skopje

petreska.anita@uklo.edu.mk, blagoj.ristevski@uklo.edu.mk, ivica.dimitrovski@gmail.com, saso.nikolovski@makpetrol.com.mk

## Abstract:

Adaptive biometric systems represent a significant advancement in authentication technologies, designed to dynamically update biometric templates to maintain high accuracy and security over time. This paper examines the key strategies for adapting these systems, focusing on their ability to evolve as users' biometric data changes due to factors like aging, environmental shifts, or behavioural variations. Central to this process are adaptation mechanisms, such as additive and replacement techniques, which continuously update biometric templates to reflect new data while maintaining the integrity of the system. Key criteria for adaptation include oracle-based methods, request acceptance, and dual thresholds, which help determine when and how the system should adapt. Adaptation modes, whether supervised or semi-supervised, guide how templates are updated based on labelled or unlabelled data. Both offline (delayed) and online (real-time) adaptation methods offer distinct advantages, depending on system requirements. Despite their benefits, adaptive biometric systems face significant challenges, particularly in large-scale applications, data acquisition, and security risks such as poisoning attacks. Addressing these challenges, including the development of secure, scalable, and resource-efficient solutions, is crucial for the widespread adoption of adaptive biometric systems, especially in mobile and commercial contexts.

## Keywords:

biometric pattern, biometric template, adaptive biometric systems, biometric databases.

## 1. Introduction

This paper aims to investigate adaptive biometric systems, which dynamically update biometric templates to improve identification accuracy and security, to improve their functionality. The introductory section explains the importance of adaptive biometric systems, focusing on their role in improving the accuracy and security of authentication [1], Section 2 covers adaptation strategies, including models for updating biometric templates and the mechanisms involved in their adaptation [2], [3], Section 3 discusses adaptation criteria, which define when and how biometric data in systems are updated [4], [5], Section 4 deals with the adaptation mode, explaining the methods of supervised and semi-supervised system adaptation. Section 5 explores adaptation mechanisms, analysing different approaches for dynamically updating biometric templates. Section 6 highlights open challenges, such as system scalability, security issues, and the need for improved test databases. The conclusion highlights that while adaptive biometric systems offer significant benefits, the challenges remain, particularly in security and adaptability.

## 2. Strategies for adaptation of the biometric template

Adaptation strategies in biometric systems are essential for improving system performance and accuracy over time. These strategies, shown in Fig. 1, are typically structured around several key components:

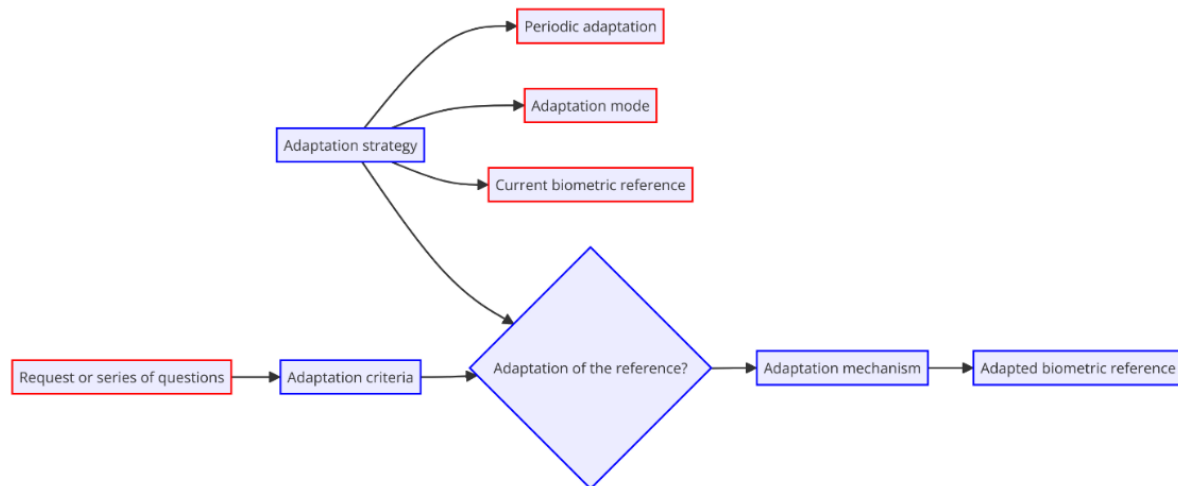


Figure 1: Generic flowchart of the biometric adaptation process.

- Reference modelling defines the structure and method for modelling the biometric template.
- Adaptation criteria determine the conditions under which the adaptation mechanism is triggered.
- Adaptation mode specifies the method used for assigning labels to biometric templates, which can be supervised or semi-supervised.
- Adaptation periodicity establishes the frequency at which the adaptation process is applied, whether in real-time (online) or periodically (offline).
- Adaptation mechanism outlines the specific steps that are taken once the adaptation criteria are met.

These strategies ensure that biometric systems can evolve and adapt to new data and environmental changes, improving their reliability and accuracy over time.

## 2.1. Reference Modelling in Adaptive Biometric Systems

Reference modelling plays a key role in adaptive biometric systems, addressing the need to adapt biometric templates over time to improve identification accuracy and reliability. During the authentication process, the user's biometric characteristics may vary due to factors such as emotional state, health conditions, or changes in the environment. These strategies are usually based on classification algorithms and biometric template update mechanisms.

In adaptive biometric systems, biometric templates are generally categorized into three main types [5]:

- A single-sample biometric template usually consists of a single sample obtained during the registration phase. Although simple, it has limitations, especially for behavioural modalities that exhibit significant variability, such as keystroke dynamics. Research shows that while this method may be sufficient for physical modalities with less variability, behavioural modalities often require multiple samples to achieve stability and accuracy [6].

- Multi-sample biometric template (galleries) involves collecting multiple samples during the registration phase and storing them in a biometric template gallery. The use of multiple samples minimizes errors, as it accounts for variations in the user's biometric data [7].

- Multiple templates per user is a method that maintains several biometric templates for each user, customizing them over time by incorporating new samples. This provides a more accurate representation of the user's current state, accommodating variations in biometric characteristics. Techniques such as clustering can automatically organize patterns into groups, with the most representative samples used as prototypes, improving system accuracy while reducing memory and processing requirements.

### 2.1.1. Adaptation Strategies

Adapting biometric systems involves various strategies that enable them to adjust to changes in a user's biometric data, as illustrated in Fig. 2. Key factors in developing these strategies include:

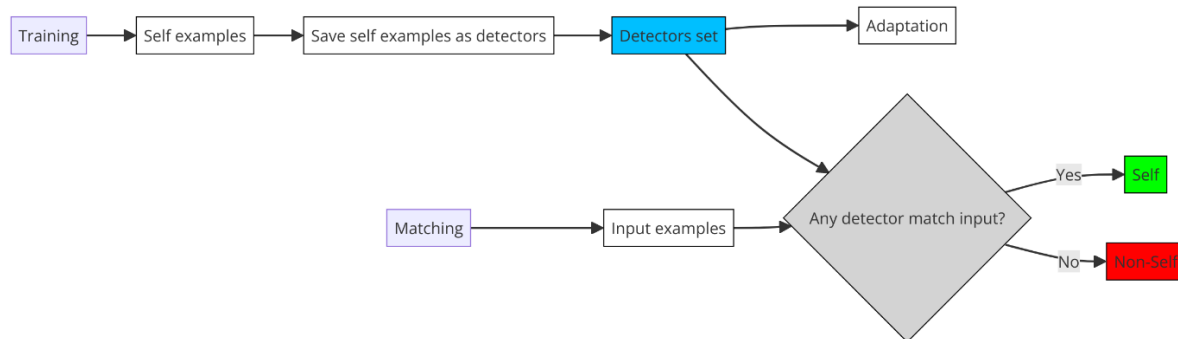


Figure 2: Algorithm for adaptive positive selection

- *Adaptation criterion*, which uses various triggers such as system performance changes or new user data to initiate adaptation.
- *Adaptation mode*, where adaptation can be either supervised or semi-supervised depending on whether the update process is automated or requires human intervention.
- *Adaptation periodicity*, where adaptation can occur either online, with each recognition event, or offline, periodically.

By improving adaptation models and developing new classification algorithms, biometric systems can significantly enhance their accuracy and reliability over time, thereby providing a higher level of protection for users.

## 3. Adaptation Criteria in Biometric Systems

Adaptation criteria are essential components that determine when and how to adapt a biometric system to maintain or improve its performance. The nature of biometric data, influenced by environmental factors, physiological changes, and operational conditions, demands continuous adjustment to maintain accuracy and security. In the literature, several adaptation criteria have been proposed, and their effectiveness varies depending on the system's needs and the biometric modality used.

*Oracle-based Adaptation:* The oracle-based approach relies on a human operator (or an equivalent decision-making entity) to decide when the system should undergo adaptation. While this method ensures high accuracy, as it relies on human judgment, it is not always feasible, especially in fully automated systems [8]. In systems requiring minimal human intervention, relying on an operator to monitor each decision is impractical. This approach is particularly useful when the stakes of incorrect classification are high, such as in critical security environments, but the operational cost can be high due to the manual intervention required.

*Request Acceptance:* The request acceptance criterion is one of the simplest forms of adaptation. In this case, any accepted authentication request is used to update the biometric template. For example, systems that use keystroke dynamics for authentication constantly monitor and update the user's typing behaviour [9]. Over time, this helps the system adapt to natural changes in a user's behaviour. However, this approach has limitations, as it risks incorporating incorrectly classified samples into the original biometric template, potentially degrading the system's accuracy.

In keystroke dynamics, for instance, changes in the typing style over time can lead to false rejections. To address this, dynamic updating of training data is recommended to ensure that classifiers remain effective even as typing behaviour evolves. There are two primary update mechanisms: "moving window" and "growing window." The "moving window" method uses a fixed number of recent samples, whereas the "growing window" incorporates all new and original samples.

*Dual Threshold Criterion:* The dual threshold approach incorporates two distinct thresholds: a decision threshold and an adaptation threshold. Only samples that meet the more stringent adaptation threshold are used to update the system [10]. This minimizes the risk of including false positives in the genuine biometric template. Although this method reduces errors associated with incorporating false samples, its improvement in overall system performance may be limited due to its strict adaptation strategy.

The dual threshold approach is beneficial in environments where it is crucial to limit errors, such as in fingerprint recognition systems, where high security and accuracy are paramount. However, its conservative nature can sometimes prevent the system from adequately adapting to genuine changes in the user's biometric data.

*Adaptive Thresholds:* The adaptive thresholds method is an enhancement of the dual threshold approach, allowing thresholds to adjust over time [11]. By continuously updating these thresholds, the system can track intra-class variations in user characteristics and adapt more effectively. This method, validated through keystroke dynamics, demonstrated improved authentication performance by reducing error rates compared to fixed thresholds.

In a biometric modality such as keystroke dynamics, where intra-class variation is significant, adaptive thresholds provide flexibility, improving both security and usability. Research has shown that adaptive thresholds reduce equal error rates (EER) by up to 2% compared to static thresholds [11].

*Quality Index:* The quality index criterion ensures that only high-quality data is incorporated into the biometric template. This index assesses the quality of biometric samples based on various factors, such as sensor performance and environmental conditions. If the data quality is deemed insufficient, the sample is discarded to prevent degrading the template's accuracy [12].

This method is particularly relevant in multimodal biometric systems, where multiple modalities (e.g., face recognition, fingerprint recognition) are used. It ensures that the system adapts only when the new data is reliable, which is crucial when the biometric system operates in uncontrolled environments.

*Condition-Sensitive Adaptation:* Condition-sensitive adaptation focuses on adjusting the system based on specific conditions during the acquisition of biometric data. For instance, if the lighting or user pose differs significantly from the reference conditions, the system adapts accordingly [13]. This approach ensures that unnecessary samples are excluded, reducing the risk of template contamination. For face recognition systems, which are highly sensitive to environmental changes, condition-sensitive adaptation can significantly improve performance by filtering out low-quality samples and focusing on conditions that resemble the reference environment.

*Error Deviation Prediction:* The error deviation prediction criterion analyses deviations in system performance to determine when adaptation is necessary. This approach is useful in scenarios where the system needs to monitor for systematic errors, such as false mismatches, which may indicate that the template is becoming outdated. Continuous monitoring of error distributions can guide decisions about when to update biometric templates.

In systems deployed at border control checkpoints, a series of false mismatches may prompt the system to initiate an update. This method helps maintain performance in dynamic, real-world environments where biometric data may evolve over time.

*Mixed Criteria:* Mixed criteria involve combining multiple sources of information to support adaptation decisions. This method is particularly effective in multimodal biometric systems, where different modalities (such as face and fingerprint recognition) can be leveraged together to improve accuracy [15]. Adaptation is triggered when one modality reaches a mismatch threshold while the other confirms the user's identity. The mixed criteria method allows for more robust adaptation strategies, especially when different biometric modalities exhibit varying degrees of intra-class variability. It has been shown to enhance template update processes and improve overall system performance [14], [15], [16].

## 4. Adaptation Mode

Adaptation criteria are essential components that determine when and how a biometric system should adapt to maintain or improve its performance. Several adaptation criteria have been proposed in the literature, with their effectiveness varying based on the system's needs and the biometric modality used.

*Oracle-based adaptation* relies on a human operator or decision-making entity to decide when the system should undergo adaptation. While this method ensures high accuracy, it is not always feasible, especially in fully automated systems. In systems requiring minimal human intervention, relying on an operator is impractical. This approach is useful in critical security environments but comes with high operational costs due to manual intervention [17].

*Request acceptance* is a simpler form of adaptation where any accepted authentication request updates the biometric template. For example, systems using keystroke dynamics monitor and adapt to changes in typing behaviour over time. However, this can introduce errors if incorrectly classified samples are incorporated. In keystroke dynamics, changes in typing style can lead to false rejections, so dynamic updating of training data through "moving window" or "growing window" mechanisms is recommended.

*The dual threshold criterion* involves two thresholds—a decision threshold and an adaptation threshold. Only samples that meet the adaptation threshold are used for updates, minimizing the inclusion of false positives but limiting system performance improvement. This method is beneficial in high-security environments like fingerprint recognition but may prevent adaptation to genuine changes [18].

*Adaptive thresholds* improve upon the dual threshold approach by allowing thresholds to adjust over time, enabling the system to track intra-class variations and adapt more effectively. Research shows that adaptive thresholds, particularly in keystroke dynamics, can reduce error rates compared to static thresholds.

*Quality index* ensures that only high-quality data is included in the biometric template by assessing biometric samples based on sensor performance and environmental conditions, discarding low-quality data to preserve template accuracy. This method is especially relevant in multimodal systems like face and fingerprint recognition, where adapting only to reliable data is crucial in uncontrolled environments.

*Condition-sensitive adaptation* adjusts the system based on specific conditions during data acquisition, such as lighting or pose changes, to ensure only relevant samples are used, improving performance in systems like face recognition that are sensitive to environmental changes.

*Error deviation prediction* analyses deviations in system performance to decide when adaptation is necessary, helping systems monitor systematic errors such as false mismatches. This method is particularly useful in dynamic environments like border control checkpoints, where biometric data can evolve over time.

*Mixed criteria* combine multiple sources of information to support adaptation decisions. This is effective in multimodal systems, where different modalities like face and fingerprint recognition can be leveraged together. Adaptation is triggered when one modality reaches a mismatch threshold while the other confirms the user's identity, enabling more robust adaptation strategies and improving overall system performance.

### 4.1. Periodicity of adaptation

The frequency of adaptation refers to how often a biometric system updates its templates. Adaptation can occur either offline (with a delay) or online (in real-time), with each approach having different effects on system performance and complexity [19].

*Offline adaptation* involves collecting biometric data over time and temporarily storing it. Once enough data is gathered or a certain time has passed, the templates are updated. This approach minimizes the impact on real-time performance since updates occur during inactive periods. The adaptation frequency is typically based on predefined intervals, making it a controlled but slower process [20]. While this method is more stable and suitable for systems that don't require immediate updates, it may struggle in environments where frequent changes occur.

*Real-time adaptation*, on the other hand, updates templates as soon as a decision criterion is met. It allows continuous template adjustments during the user verification process, helping maintain high accuracy as conditions change. However, this adds complexity and processing time because recognition and adaptation happen simultaneously, making it more suitable for environments where biometric data or user behaviours frequently change [21].

The choice between offline and real-time adaptation depends on the system's requirements, resource availability, and desired accuracy. Offline adaptation helps maintain stable system performance and avoids overburdening resources during operation, though it may not be responsive enough for environments that require frequent updates. Real-time adaptation offers immediate responsiveness to changing conditions, but this comes at the cost of added complexity and processing delays. It is particularly useful in critical applications like security and access control, where biometric data and environmental conditions change frequently. However, for systems where stability and minimal processing overhead are prioritized, offline adaptation remains a strong option, allowing periodic updates without disrupting system performance.

## 5. Adaptation and Additive Mechanism in Biometric Systems

The adaptation mechanism in biometric systems is crucial to maintaining high performance in dynamic environments. As biometric data, such as fingerprints, facial recognition or typing dynamics, evolve over time due to factors such as ageing, changes in lighting or variations in user behaviour, the system needs to update its biometric templates to prevent accuracy degradation. The adaptation process ensures that the biometric system remains resilient and accurate despite these variations. The four basic categories of adaptation mechanisms are presented: additive mechanisms, substitution mechanisms, multi-gallery mechanisms, and selection mechanisms [22].

Additive mechanisms incrementally add new samples to the biometric gallery without removing existing ones. These mechanisms help capture a wider range of within-class variation in a user's biometrics, reducing false rejections that occur when genuine biometric variations are confused with imposter attempts. The augment-update mechanism is one of the earliest studied methods, as explored in studies such as [23], which continuously adds new, original samples to the user's gallery. This method assumes that the new samples are reliable and represent the true variations in the user's biometric data.

The advantage of additive mechanisms lies in their simplicity and ability to handle a wide range of biometric variations. However, a significant drawback is the uncontrolled growth of the biometric gallery over time, which can lead to storage and computational inefficiencies. Mechanisms like Graph Min-Cut use advanced algorithms to add new samples while retaining basic information about the user's biometric patterns without overloading the system's memory [24].

Replacement mechanisms not only add new samples to the gallery, but also remove older or less representative samples. This strategy addresses the limitation of additive mechanisms by preventing the gallery from growing indefinitely. A sliding window mechanism, also known as "First In First Out" (FIFO), replaces the oldest samples in the gallery with newer ones. This keeps the size of the gallery constant, making it manageable and preventing memory overload [25].

More advanced replacement mechanisms, such as MDIST and DEND, prioritize the replacement of samples based on their similarity to other samples in the gallery. MDIST selects samples that are most representative of the user's biometric data by calculating their average distance to other samples in the gallery, while DEND focuses on preserving samples that capture a wide range of variation within the class. Both methods have been shown to improve performance in biometric systems, especially when used in combination with semi-supervised learning techniques [26]. Another innovative replacement mechanism is Least Recently Used (LRU), which replaces the least recently used samples in the gallery. Although efficient in some scenarios, LRU can be computationally expensive, as it requires the system to track the usage frequency of each sample in the gallery [27].

Multi-gallery mechanisms involve managing two or more biometric galleries simultaneously, with each gallery using a different adjustment mechanism. This approach allows biometric systems to benefit from the strengths of multiple adaptation strategies. For example, in a dual parallel configuration, one gallery can use the Growing Window mechanism to save initial user templates, while another gallery can use the Sliding Window to update the system with the latest biometric data. The combined results

of the two galleries improve the system's ability to adapt to user behaviour while maintaining pattern integrity.

Another example is Co-Update, a method that applies the principles of co-training to adaptive biometric systems. In this approach, each gallery corresponds to a different biometric modality, such as face and fingerprint recognition. When one modality confidently classifies a sample, the corresponding gallery for the other modality is updated. This approach leverages the complementary nature of different biometric traits to improve system accuracy and robustness in real-world scenarios [28].

Selection mechanisms aim to reduce the size of the biometric gallery by retaining only the most representative samples. These mechanisms are often used in conjunction with additive mechanisms to manage gallery growth effectively. In the context of template selection, the system identifies the most relevant samples that accurately represent the user's biometric patterns. For example, the DEND algorithm uses hierarchical clustering to group samples based on their similarity and selects the most representative sample from each cluster for inclusion in the gallery [29].

An alternative approach is MDIST, which ranks samples based on their average similarity to other samples in the gallery. Samples with the highest similarity scores are retained, while others are discarded. Selection mechanisms are particularly useful in scenarios where storage limitations or computational efficiency is a concern, as they reduce the overall size of the biometric gallery without sacrificing accuracy [30].

## 6. Open Challenges

The development of adaptive biometric systems brings numerous challenges, especially in scaling, dataset acquisition, and defence against attacks. These systems must constantly adapt to changes in biometric data over time, such as ageing or environmental factors, to maintain high accuracy and security. Below are some of the key open challenges in adaptive biometric systems [31], [32], [33], [34].

*Large-Scale Adaptive Biometric Systems:* Most current studies on adaptive biometric systems focus on small to medium-sized datasets with around 33 to 550 users [29]. However, in real-world applications like border control, the number of users can be significantly larger. Scaling these systems raises concerns regarding the time and resources required for adaptation. While small datasets are useful for proof of concept, larger datasets are crucial for evaluating the true capabilities and limitations of adaptive systems. Future research needs to explore how large-scale datasets affect adaptation strategies in real-time systems.

*Acquisition of Suitable Datasets:* One of the most pressing issues in the field is the lack of large and varied biometric datasets that can be used to evaluate adaptive biometric systems. To properly assess the adaptability of these systems, datasets should include multiple samples per user, ideally collected over different sessions under varying conditions. However, publicly available datasets often lack such variety, making it challenging to evaluate the performance of systems in real-world settings. Gathering long-term data for adaptive systems is time-consuming, but it is critical to improving the accuracy and resilience of biometric recognition technologies.

*Generation of Synthetic Datasets:* Since collecting real biometric data over long periods is difficult, generating synthetic datasets has become a popular alternative. These datasets can simulate temporal changes in biometric features, allowing researchers to test algorithms on large-scale data. Recent work has explored this approach for various modalities, including fingerprint and keystroke dynamics. However, challenges remain in accurately simulating real-world biometric variations, especially as changes differ across modalities. Future research should focus on modelling how temporal changes affect different biometric features to enhance synthetic dataset realism.

*Attacks on Biometric Systems:* A significant threat to adaptive biometric systems comes from attacks such as spoofing, replay, and poisoning attacks. Poisoning attacks occur when adversaries intentionally introduce false data into the system, leading to the corruption of genuine biometric templates. Adaptive systems, which update their templates over time, are particularly vulnerable to such attacks because they can mistakenly incorporate malicious data as legitimate input. Countermeasures like liveness detection and ensemble techniques can help mitigate these vulnerabilities. Nonetheless, adaptive



biometric systems must balance between adaptability and security to avoid being exploited by adversaries.

*Adaptation to Both Temporal and Environmental Conditions:* Biometric systems must adapt not only to temporal changes, such as ageing but also to varying acquisition conditions. For example, in facial recognition systems, templates may need to adapt to different lighting or facial poses. Systems like Self-Update can adapt over time, while others like Sliding Window are better suited for adapting to changing conditions. A key challenge is designing systems that can adapt to both types of changes without compromising performance.

*Commercial Applications of Adaptive Biometric Systems:* While the potential benefits of adaptive biometric systems are well-documented, few commercial applications have implemented these strategies. For instance, Apple's Face ID uses adaptive methods to improve face data over time, but many other systems remain static. One of the major challenges in commercial adoption is the vulnerability of adaptive systems to attacks, which could undermine user trust. To encourage wider adoption, developers must implement robust security measures to prevent attackers from corrupting biometric templates during the adaptation process.

*Adaptive Biometrics in Mobile Applications:* Mobile devices present unique challenges for biometric systems due to limited computational resources and storage. However, by offloading processing tasks to cloud servers, adaptive biometric systems could be effectively implemented on mobile platforms. Ensuring a secure communication channel for transmitting biometric data between mobile devices and cloud servers is essential. Future research should also focus on gathering long-term sensor data from mobile devices to improve adaptive biometric models on mobile platforms.

## 7. Conclusion

Adaptive biometric systems represent a significant step forward in authentication technology, allowing dynamic adaptation of biometric templates according to changes in users and their working conditions. Unlike traditional biometric systems, whose performance degrades over time due to natural changes in users, adaptive systems offer mechanisms to continuously update templates. This allows the systems to maintain high accuracy and security, even in complex scenarios.

One of the key benefits of these systems is their ability to minimize mismatch errors caused by user variations, such as ageing, posing changes, or acquisition conditions. This improves system accuracy and reduces the need for users to re-enrol, which in turn lowers maintenance costs. Additionally, adaptive biometric systems provide flexibility, which is particularly important for commercial applications where acquisition conditions often change.

However, despite the many advantages, these systems face significant challenges, particularly regarding security and privacy. One of the major threats is poisoning attacks, where malicious users introduce false data into the system, potentially compromising the biometric templates. Since adaptive systems continuously update their templates, they are particularly vulnerable to such attacks. This calls for the development of new protective measures, such as liveness detection mechanisms or the use of multimodal systems to enhance security.

Another open challenge in this field is managing large-scale datasets and system adaptation to them. Most research so far has focused on small user groups, while real-world applications, such as border control, require systems capable of handling thousands of users. In the future, there will be a need to develop systems that can adapt to large-scale datasets while maintaining efficiency.

A significant challenge is also the collection of appropriate datasets for evaluating adaptive biometric systems. The datasets need to include multiple samples per user, collected across different sessions and conditions, to allow for realistic system evaluations. However, such data is difficult to obtain, and collecting it is expensive and time-consuming. As an alternative, generating synthetic datasets is being considered, allowing for testing adaptive algorithms in simulated conditions. This approach is promising, but challenges remain in accurately simulating real-world biometric changes.

In the future, research should focus on the development of adaptive systems for mobile devices, where computational and storage resources are limited. Offloading data processing to the cloud is one way to address this issue, though secure communication channels for transmitting biometric data

between mobile devices and cloud servers remain a challenge. Moreover, long-term data collection from mobile sensors is crucial for improving adaptive models in mobile platforms.

In conclusion, adaptive biometric systems are a new and significant technology offering solutions to many limitations of traditional systems. However, to enable broader adoption, current challenges related to security, data management, and adaptation to new conditions must be addressed.

#### References:

- [1] Ryu, R., Yeom, S., Herbert, D., & Dermoudy, J. (2023). The design and evaluation of adaptive biometric authentication systems: Current status, challenges and future direction. *ICT Express*, 9(6), 1183-1197.
- [2] Zhang, X., Ye, H., Wang, J., Lin, Y., Xiong, J., Hwu, W. M., & Chen, D. (2020, November). DNNExplorer: a framework for modelling and exploring a novel paradigm of FPGA-based DNN accelerator. In *Proceedings of the 39th International Conference on Computer-Aided Design* (pp. 1-9).
- [3] Gupta, K., Walia, G. S., & Sharma, K. (2020). Quality based adaptive score fusion approach for multimodal biometric system. *Applied Intelligence*, 50(4), 1086-1099.
- [4] Kortli, Y., Jridi, M., Al Falou, A., & Atri, M. (2020). Face recognition systems: A survey. *Sensors*, 20(2), 342.
- [5] N. J. Grabham and N. M. White, "Use of a novel keypad biometric for enhanced user identity verification," *In Proceedings of the Instrumentation and Measurement Technology Conference (IMTC'08)*, 2008.
- [6] C. Pagano, E. Granger, R. Sabourin, P. Tueri, G. L. Marcialis, F. Roli, "Context-sensitive self-updating for adaptive face recognition," *Adaptive Biometric Systems*. Springer, p. 9–34, 2015.
- [7] Alessandra Lumini and Loris Nanni, "A clustering method for automatic biometric template selection," *tom* 39, 6p. 3, p. 495–497, 2006.
- [8] Rahul Sukthankar and Robert Stockton, "Argus: The digital doorman," *tom* 16, 6p. 2, pp. 14-19, 2001.
- [9] Pilsung Kang, Seong-seob Hwang, Sungzoon Cho, "Continual retraining of keystroke dynamics-based authenticator," *In Advances in Biometrics*. Springer, p. 1203–1211, 2007
- [10] Ajita Rattani, "Adaptive biometric system based on template update procedures,," PhD Thesis, Department of Electrical and Computer Engineering, University of Cagliari, 2010.
- [11] Abir Mhenni, Christophe Rosenberger, Estelle Cherrier, Najoua Essoukri Ben Amara, "Keystroke template update with adapted thresholds," *bo%1 Proceedings of the International Conference on Advanced Technologies for Signal and Image Processing*, 2016.
- [12] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, A. K. Jain, "Second fingerprint verification competition," *bo%1 Proceedings of the 16th International Conference on Pattern Recognition*, 2002.
- [13] C. Pagano, E. Granger, R. Sabourin, P. Tueri, G. L. Marcialis, F. Roli, "Context-sensitive self-updating for adaptive face recognition," *Adaptive Biometric Systems*. Springer, p. 9–34, 2015.
- [14] Aleix Martinez, Robert Benavente, "The AR Face Database,," CVC Technical Report 24. Centre de Visió per, 1998.
- [15] Fabio Roli, Luca Didaci, and Gian Luca Marcialis, "Template co-update in multimodal biometric systems," *Advances in Biometrics*. Springer, p. 1194–1202, 2007.
- [16] Norman Poh, Amin Merati, Joseph Kittler, "Adaptive client-impostor centric score normalization: A case study in fingerprint verification," *bo%1 3rd International Conf. on Biometrics: Theory, Applications, and Systems*, 2009.
- [17] Paulo Henrique Pisani, Norman Poh, André C. P. L. F. de Carvalho, Ana Carolina Lorena, "Score normalization applied to adaptive biometric systems," *Comput. Secur.*, *tom* 70, p. 565–580, 2017.
- [18] ZiboWang, Abdul Serwadda, Kiran S. Balagani, and Vir V. Phoha, "Transforming animals in a cyber-behavioral biometric menagerie with frog-boiling attacks," *bo%1 5th International Conference on Biometrics: Theory, Applications and Theory*, 2012.
- [19] Paulo Henrique Pisani, Ana Carolina Lorena, and André C. P. L. F. de Carvalho, "Adaptive algorithms in accelerometer biometrics," *bo%1 Brazilian Conference on Intelligent Systems (BRACIS'14)*, 336–341, 2014.

- [20] Indrė Žliobaitė, Albert Bifet, Jesse Read, Bernhard Pfahringer, Geoff Holmes, "Evaluation methods and decision theory for classification of streaming data with temporal dependence.," *Mach. Learn.*, Vol. 98, 6p. 3, p. 455–482, 2015.
- [21] Umut Uludag, Arun Ross, and Anil Jain, "Umut Uludag, Arun Ross, and Anil Jain," *Biometric template selection and update: A case study in fingerprints*, 2004.
- [22] Ajita Rattani, Gian Luca Marcialis, and Fabio Roli, "Self adaptive systems: An experimental analysis of the performance over time," *Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM'11)*. 36–43., 2011
- [23] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training," *Workshop on Computational Learning Theory*. 92–100., 1998.
- [24] Romain Giot, Mohamad El-Abed, Baptiste Hemery, Christophe Rosenberger, "Unconstrained keystroke dynamics authentication with shared secret.," *Comput. Secur.*, том 30, 6p. 6, p. 427–445, 2011.
- [25] Paulo Henrique Pisani, Ana Carolina Lorena, and André C. P. L. F. de Carvalho, "Adaptive algorithms in accelerometer biometrics," *Brazilian Conference on Intelligent Systems (BRACIS'14)*, 336–341, 2014.
- [26] Indrė Žliobaitė, Albert Bifet, Jesse Read, Bernhard Pfahringer, Geoff Holmes, "Evaluation methods and decision theory for classification of streaming data with temporal dependence.," *Mach. Learn.*, том 98, 6p. 3, p. 455–482, 2015.
- [27] Umut Uludag, Arun Ross, and Anil Jain, "Umut Uludag, Arun Ross, and Anil Jain," *Biometric template selection and update: A case study in fingerprints*, 2004. Ajita Rattani, Gian Luca Marcialis, and Fabio Roli, "Self adaptive systems: An experimental analysis of the performance over time," *Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM'11)*. 36–43., 2011.
- [28] Romain Giot, Mohamad El-Abed, Baptiste Hemery, Christophe Rosenberger, "Unconstrained keystroke dynamics authentication with shared secret.," *Comput. Secur.*, том 30, 6p. 6, p. 427–445, 2011.
- [29] Pisani, P. H., Mhenni, A., Giot, R., Cherrier, E., Poh, N., Ferreira de Carvalho, A. C. P. D. L., ... & Amara, N. E. B. (2019). Adaptive biometric systems: Review and perspectives. *ACM Computing Surveys (CSUR)*, 52(5), 1-38.
- [30] Ajish, S., & AnilKumar, K. S. (2023). Performance enhancement of symmetric hashed fingerprint template using dynamic threshold matching algorithm. *International Journal of Biometrics*, 15(1), 78-100.
- [31] Vensila, C., & Boyed Wesley, A. (2024). Multimodal biometrics authentication using extreme learning machine with feature reduction by adaptive particle swarm optimization. *The Visual Computer*, 40(3), 1383-1394.
- [32] Abdullahi, S. M., Sun, S., Wang, B., Wei, N., & Wang, H. (2024). Biometric template attacks and recent protection mechanisms: A survey. *Information Fusion*, 103, 102144.
- [33] Harikrishnan, D., Sunil Kumar, N., Joseph, S., & Nair, K. K. (2024). Towards a fast and secure fingerprint authentication system based on a novel encoding scheme. *International Journal of Electrical Engineering & Education*, 61(1), 100-112.
- [34] Gupta, K., Walia, G. S., & Sharma, K. (2020). Quality based adaptive score fusion approach for multimodal biometric system. *Applied Intelligence*, 50(4), 1086-1099.

# Using Machine Learning Algorithms of Stroke Prediction

Daniela Slavkovska<sup>1</sup>, Anita Petreska<sup>2</sup>, Blagoj Ristevski<sup>2</sup>, Saso Nikolovski<sup>4</sup>.<sup>1</sup>P knqr: "T gpf gxunk<sup>4</sup>

<sup>1</sup> Children's Hospital Skopje

<sup>2</sup> Faculty of Information and Communication Technologies - Bitola, University "St. Kliment Ohridski" - Bitola, Republic of North Macedonia

dslavkovska@yahoo.com; blagoj.ristevski@uklo.edu.mk; etreska.anita@uklo.edu.mk;  
sasnik@gmail.com; nikola.rendevski@uklo.edu.mk

## Abstract:

Stroke is a severe medical condition resulting from disrupted blood flow or ruptured blood vessels in the brain, often leading to life-threatening consequences. The World Health Organization (WHO) identifies stroke as a leading cause of death and disability worldwide. Although significant research has focused on heart-related diseases, stroke prediction has received comparatively less attention. To address this gap, this paper presents machine learning models developed to predict stroke likelihood, utilizing key physiological factors associated with stroke risk. Six algorithms: logistic regression, decision tree, random forest, KNN, SVM and Naïve Bayes, were implemented to train and test prediction models. The primary objective was to determine the algorithm that provides the highest predictive accuracy.

Our findings reveal that the Naïve Bayes algorithm performed best, achieving an accuracy of approximately 82%. This is notable given Naïve Bayes' suitability for probabilistic data and its efficiency in handling complex variable interactions, suggesting its value for early stroke detection in clinical settings. The use of machine learning in stroke prediction highlights a promising approach for early intervention, potentially aiding in reducing stroke-related mortality and morbidity.

This paper contributes to expanding the application of machine learning in healthcare, emphasizing the need for focused stroke prediction research. Future work could enhance these models by integrating diverse datasets, testing additional machine learning techniques, and refining predictive algorithms to boost accuracy and reliability. By advancing stroke prediction, machine learning may play a key role in mitigating stroke's impact on global health.

## Keywords:

Machine Learning, Logistic Regression, Decision Tree Classification, Random Forest Classification, KNN, SVM and Naïve Bayes

## 1. Introduction

About 11% of all deaths worldwide are due to stroke **Error! Reference source not found.**, according to the Centers for Disease Control and Prevention, strokes occur in the United States each year in about 795,000 people.

With the advancement of medical technology, machine learning can now be used to predict stroke. It is possible to make accurate predictions and analyses using machine learning algorithms. Strokes can be predicted using machine learning algorithms.

Six different machine learning algorithms were tested, with Naive Bayes achieving the highest accuracy.

Machine learning algorithms are useful for making accurate predictions and delivering accurate analytics. Research conducted on stroke has mainly focused on predicting heart attacks. The main elements of the methods used and the results achieved show that of the five classification algorithms tested, Naïve Bayes showed the highest performance by achieving a superior accuracy metric. The limitation of this model lies in its training on textual data instead of actual real-time brain images. This paper shows how six machine learning algorithms were put into practice. This paper has the potential to be extended to include the implementation of all existing machine learning algorithms.

A database from Kaggle, which contains a range of physiological traits as its attributes, was used for this paper.

These features are subsequently examined and used for final forecasting. The data is first cleaned and prepared for machine learning model understanding. Data preprocessing is the term used for this stage. To do this, the database is examined for all missing values and then filled. Then the string values are converted to integers using Label encoding, if necessary one-time encoding is performed.

After data preprocessing, the database is divided into training and testing data. Next, various classification algorithms are used to construct a model using the updated data. Accuracy is calculated for each of these algorithms and then compared to determine the most effective model for making predictions.

## 2. Machine learning algorithms for stroke prediction

Machine learning has been effectively applied to predict various diseases, such as diabetes, heart disease, and numerous efforts have been made to develop stroke prediction models using various classification techniques. Six classification algorithms were tested: Naïve Bayes, SVM, K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT) and Logistic Regression (LR). The Naïve Bayes classifier achieved the best performance, with accuracy, precision, recall and F1-score of 82%, 79.2%, 85.7% and 82.3%. Also Naïve Bayes showed the highest area under the ROC curve (AUC), reaching 82%.

## 3. Methodology

A dataset from Kaggle was used, from all available datasets the appropriate one was selected to construct the model.

### 3.1. Methodology

Once the data is collected the next task is to organize it in order to improve its clarity and ensure that it is easily interpreted by machine learning algorithms. The process is known as data preprocessing.

The process includes handling missing values, managing unbalanced data, and implementing label coding that is unique to this data set.

With the pre-processed data, they are ready to construct the model. Data-driven and machine learning algorithms are essential for model creation. Logistic regression, decision tree, random forest, K-nearest neighbor, support vector machine and Naïve Bayes classifiers were used. After constructing six different models, they are evaluated using five metrics to determine their performance: accuracy score, precision score, recall score, F1 score, and receiver operating characteristic (ROC) curve.

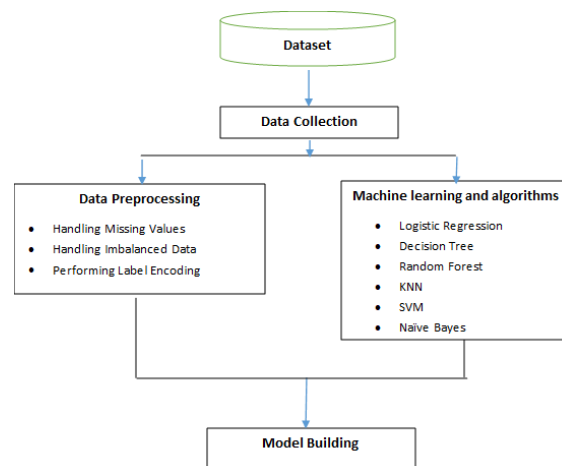


Figure 1: Methodology flow diagram

## 4. Implementation

The implementation of this paper is as follows:

- Data collection
- Data Preprocessing
- Label Encoding
- Handling Imbalanced Data

### 4.1. Data collection

The dataset used for stroke prediction was retrieved from Kaggle 0. It consists of 5110 rows and 12 columns, with the following attributes: "id", "sex", "age", "hypertension", "heart\_disease", "ever\_married", "job\_type", "stay\_type", "average\_glucose\_level", "bmi", "smoking\_status" and "stroke". The 'stroke' column serves as an output variable, with values of '0' or '1'. A value of "0" indicates no detected risk of stroke, while "1" suggests a potential risk of stroke. This database is very unbalanced, as there are significantly more occurrences of "0" (4861 rows) compared to "1" (249 rows). To improve the accuracy, pre-processing of the data is undertaken to resolve the imbalance.

**Table 1:**  
Stroke Dataset

Attribute name	Type (values)	Description
ID	Integer	A unique integer value for patients
Gender	Male, Female, Other	Tell the gender of the patient
Age	Integer	Age of patient
Hypertension	Integer (1 ,0)	Tall whether the patient has hypertension or not
Heart disease	Integer (1 ,0)	Tall whether the patient has heart disease or not
Ever married	String literal (Yes ,No)	Tall whether the patient is married or not
Work type	String literal (children, govt job, never worked, private, self- employed)	It gives different categories for work
Residence type	String literal (Urban ,Rural)	The patients residence type is stored
Avg glucose level	Floating point number	Gives the value of average glucose level in blood
Bmi	Floating point number	Gives the value of patients Body Mass Index
Smoking status	String literal (formerly, smoked, never smoked, smokes, unknown)	Gives the smoking status of the patient
Stroke	Integer (1 ,0)	A unique integer value for patients

## 4.2. Data Preprocessing

Data pre-processing is a crucial step before building a model, it helps to eliminate unwanted noise and outliers in the data set, which can hinder model performance. This phase ensures that any factors that reduce the effectiveness of the model are addressed. Once the database is collected, the next step involves cleaning the data and preparing it for model development. The database used in this case contains 12 attributes, as shown in Table 1. Initially, the "id" column is discarded, as it does not contribute to the performance of the model. The data set is then examined for missing values. For example, the column 'bmi' contains null values, which are replaced by the mean value of the column. After addressing the missing data, the next step is to perform label coding.

## 4.3. Label encoding

Label encoding transforms categorical string values into numeric values to make the data set understandable to machine learning algorithms, which typically work with numeric data. In the dataset five columns contain string values. Label encoding is applied to convert these strings to integers with a database composed entirely of numeric values.

## 4.4. Handling Imbalanced Data

The data used to predict stroke is an unbalanced data set. Out of 5110 rows only 249 indicate the occurrence of a stroke while 4861 indicate no stroke. This imbalance is illustrated in Figure 2. Training a model with such skewed data may yield high accuracy, but may result in poor precision. To solve this problem and improve the performance of the model, the data is balanced using the undersampling technique. The sample adjusts the dataset by reducing the majority class to match the size of the minority class. In this case rows with a value of "0" are undersampled to equal 249 rows with a value of "1".

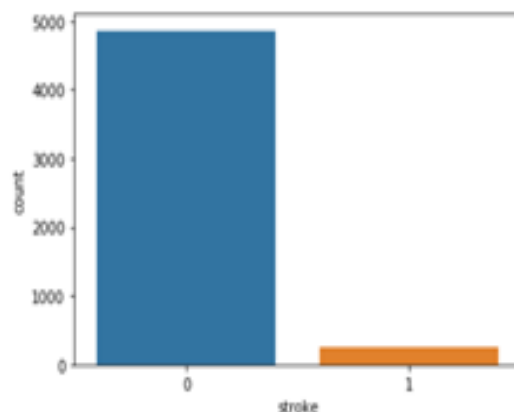


Figure 2: Before Undersampling

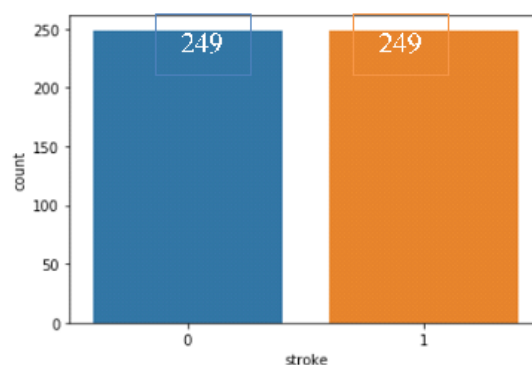


Figure 3: After Undersampling

## 4.5. Matrix correlation

The matrix correlation highlights the influence of different characteristics on an attribute. Figure 4 illustrates the relationship between other attributes and the stroke attribute. The graph shows that no single parameter has a dominant effect on the occurrence of stroke. The key factors that significantly affect the risk of stroke are: gender, age, hypertension, heart disease, average blood sugar levels, body mass index and smoking status. Attributes with the least impact are type of work, type of residence and marital status.

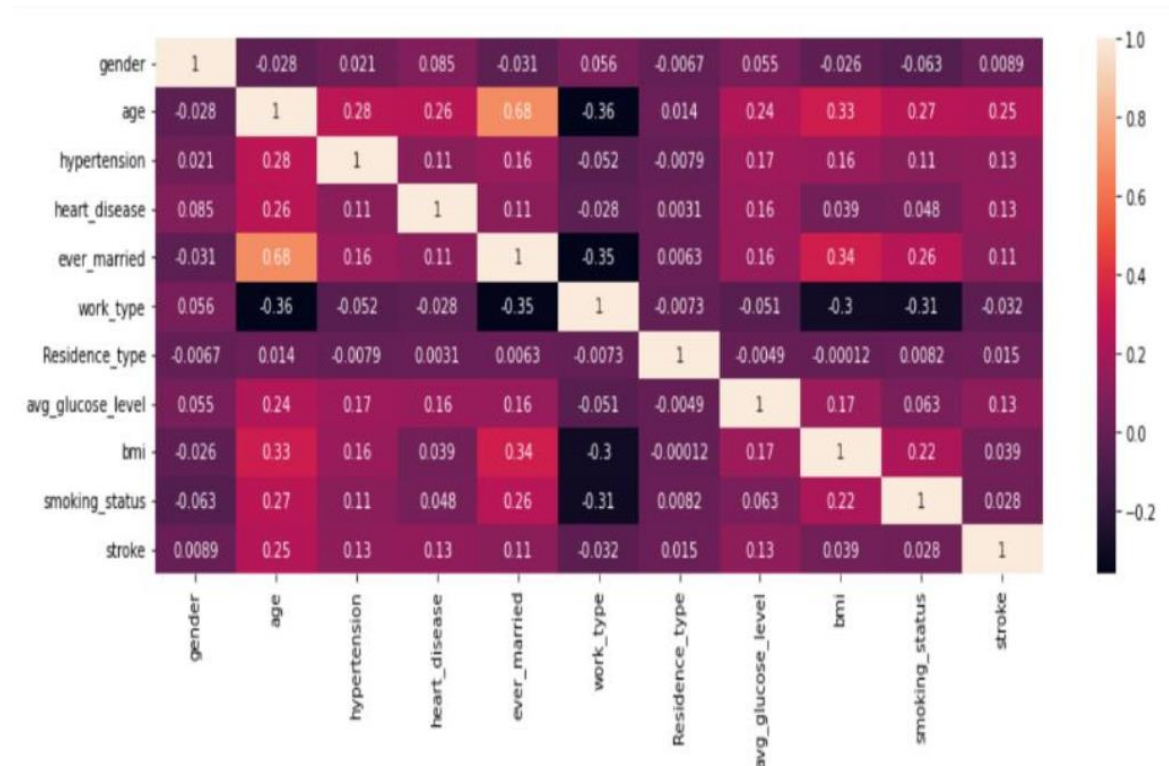


Figure 4: Matrix correlations between sociodemographics, lifestyle, and disease

## 5. Model building

### 5.1. Splitting the Data

After the data preprocessing is completed and the imbalanced database is resolved, the next step is model building. The data is split into a training dataset and a testing dataset to ensure better accuracy and efficiency. The data is split 80:20, with 80% used for training and 20% for testing. After splitting the data, several classification algorithms are applied to train the model. Algorithms used include logistic regression, decision tree classification, random forest classification, K-nearest neighbors (KNN), support vector machine (SVM), and Naïve Bayes classification.

### 5.2. Classification Algorithms

#### 5.2.1. Logistic regression

Logistic regression 0 is a supervised learning algorithm commonly used to predict the probability of a binary output variable (0 or 1). Since the output attribute in this dataset is binary, logistic regression is an ideal choice. After applying this algorithm, the model achieved an accuracy of 78%. The



performance of the algorithm will be further evaluated using additional metrics such as precision and recall, both yielded a result of 77.6%. The F1 score is also 77.6%. A receiver operating characteristic (ROC) curve for logistic regression 0 has a performance of 78%, as illustrated in Figure 5.

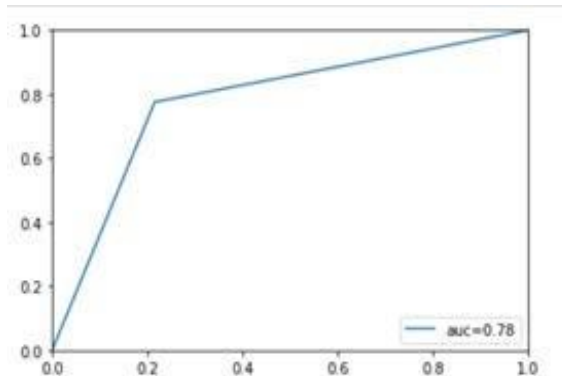


Figure 5: ROC curve for logistic regression

### 5.2.2. Decision tree classification

Decision tree 0 classification is a supervised learning algorithm used for both regression and classification tasks. It works by partitioning data based on specific parameters into a tree-like structure consisting of decision nodes (where the data is partitioned) and leaf nodes (which provide the result). In this stroke prediction model the Decision Tree algorithm achieved an accuracy of 66%, which is lower than the accuracy obtained with logistic regression. Like the logistic regression, the precision and recall scores are equal to 77.6%. The F1 score also matches this value with 77.6%. The ROC curve for decision tree classification 0 showed an accuracy of 66%, as shown in Figure 6.

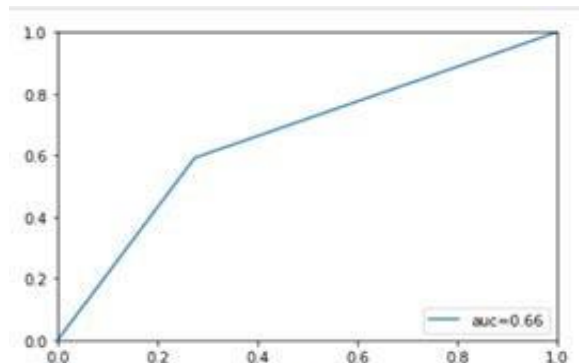


Figure 6: ROC curve for Decision Tree

### 5.2.3. Random Forest Classification

The next classification algorithm used is Random Forest Classification 0. A random forest is composed of multiple decision trees, each trained independently on random subsets of the data. During training, these trees are generated and each one gives a result. For the final prediction, a process called 'voting' is used, where each tree votes for an output class ('stroke' or 'no stroke'). The class with the most votes is chosen as the final prediction. The model achieved an accuracy of 73% using this algorithm. The precision and recall scores are 72% and 73.5%, the F1 score is 72.7%. A receiver operating characteristic (ROC) curve for random forest classification0 showed an accuracy of 73%, as shown in Figure 7.

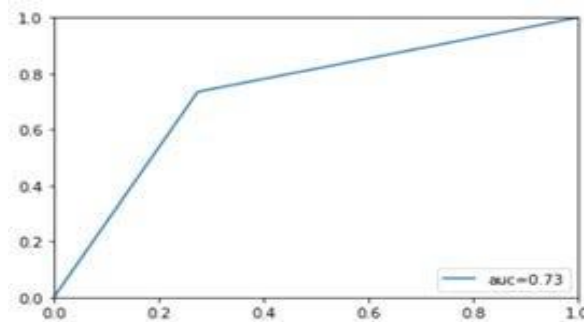


Figure 7: ROC curve for Random Forest

#### 5.2.4. K-Nearest Neighbours Classification

Another classification algorithm applied is K-Nearest Neighbors (KNN) 0, a supervised learning technique. KNN is a lazy algorithm, meaning that it does not perform training immediately after obtaining the dataset. Instead, it stores the database and acts on it only during classification. The algorithm works by finding similarities between new data and existing data, and then assigning the new data to the category most similar to the existing categories. This algorithm achieved an accuracy of 80%. Precision and recall scores were 77.4% and 83.7%, respectively. The F1 score was 80.4%. The ROC curve for KNN 0 showed an accuracy of 80%, as illustrated in Figure 8.

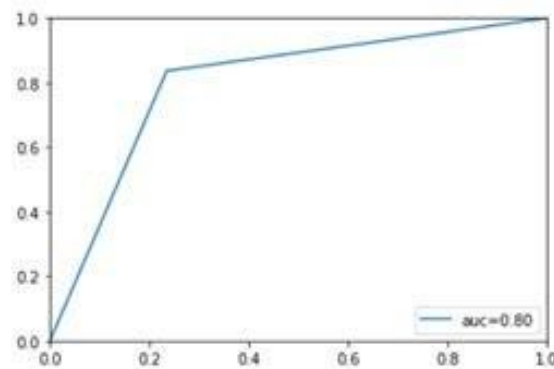


Figure 8: ROC curve for KNN

#### 5.2.5. Support Vector Machine (SVM)

Support Vector Machine 0 is a supervised learning algorithm used for both classification and regression tasks. SVM is particularly effective for high-dimensional data. In this case, the algorithm achieved an accuracy of 80%. The precision and recall scores were 78.6% and 83.8%, respectively, resulting in an F1 score of 81.1%. A receiver operating characteristic (ROC) curve for SVM 0 indicates a performance of 80%, as shown in Figure 9.

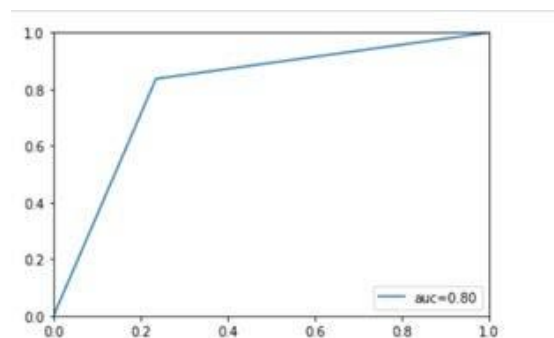


Figure 9: ROC curve for SVM.

### 5.2.6. Naïve Bayes classification

Naïve Bayes 0 is a supervised learning technique that works on the assumption that each feature is independent of the others, based on Bayes' theorem. The Naïve Bayes classifier assumes that the presence of any feature in a class is unrelated to the presence of other features. This algorithm achieved an accuracy of 82%. Precision and recall scores were 79.2% and 85.7%, respectively, with an F1 score of 82.3%. The ROC curve for Naïve Bayes 0 classification showed a performance of 82%, as shown in Figure 10.

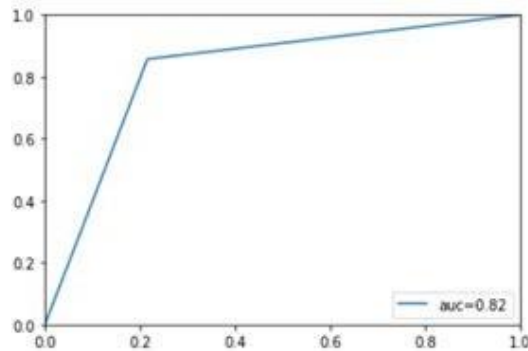


Figure 10: ROC curve for Naïve Bayes classification

After building the models, it is concluded that Naïve Bayes outperformed the other algorithms.

## 6. Conclusion

A stroke is a serious medical condition that requires prompt intervention to prevent further complications. The development of a machine learning model could aid in the early detection of stroke, helping to mitigate its long-term effects. This paper evaluates the performance of several machine learning algorithms in stroke prediction based on various physiological attributes. Among the algorithms tested, Naïve Bayes Classification performed best, achieving an accuracy of 82%. The comparison of accuracy of different algorithms is illustrated in Figure 11. In terms of precision, recall and F1 scores, Naïve Bayes also outperformed the others. These comparisons are presented in Figure 12, Figure 13 and Figure 14.

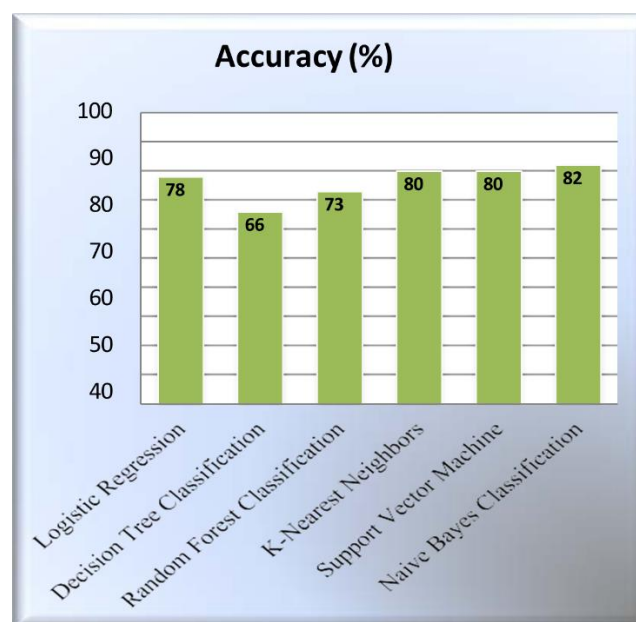


Figure 11: Comparing the Accuracies of ML Algorithms

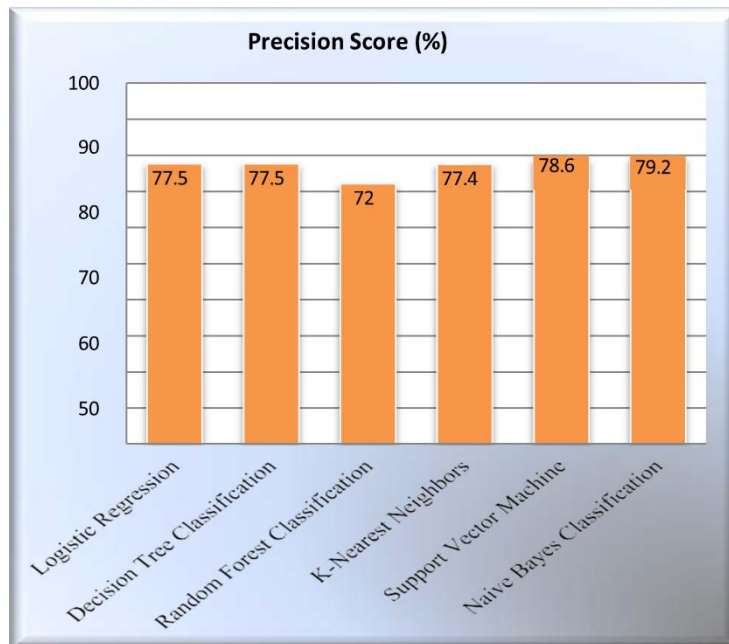


Figure 12: Comparing the Precision Scores of ML Algorithms

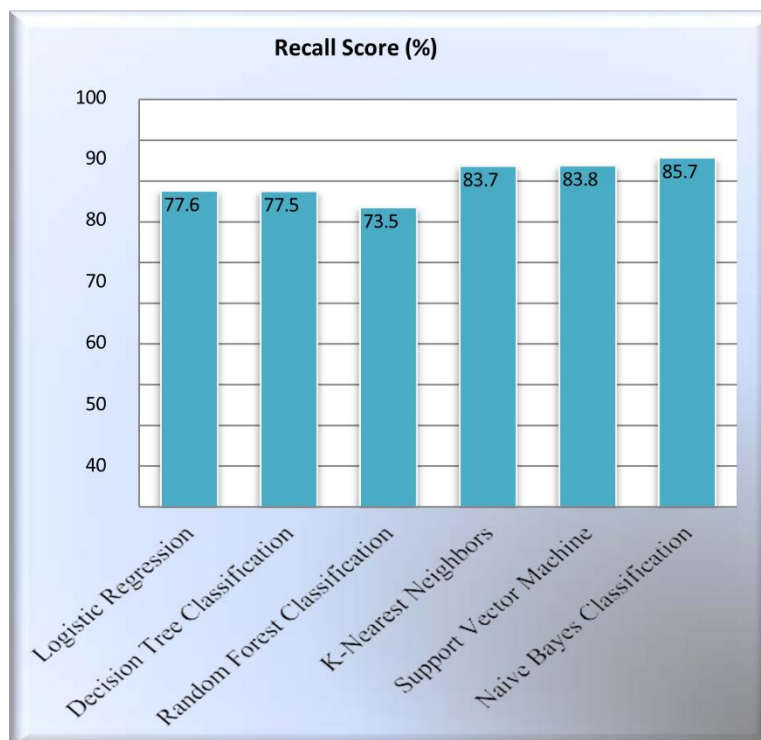


Figure 13: Comparing the Recall Scores of ML Algorithms

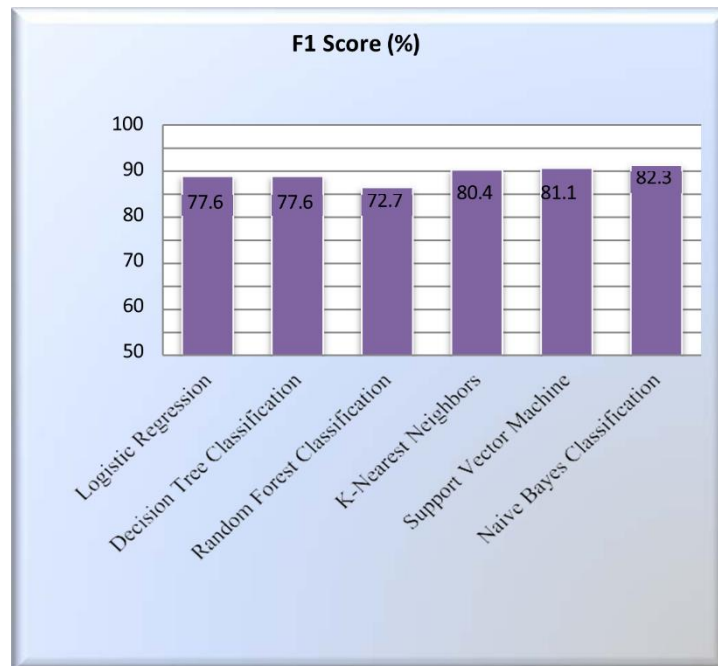


Figure 14: Comparing the F1 Scores of ML Algorithms

This paper investigates the implementation of various machine learning algorithms on the given dataset. Future work could extend this project by incorporating neural networks to train the model, allowing for a more comprehensive performance comparison using additional accuracy metrics. Currently, this study focuses on textual data, which may not always provide the most accurate predictions of stroke. A more effective approach could involve using datasets that include medical images, such as CT scans of the brain, to improve the accuracy of predicting stroke in the future.

#### References:

- [1] Olajide, A. O. "Life after stroke: more than a survival of the fittest." (2021).
- [2] Nugroho, Syamsul. *PERBANDINGAN METODE FUZZY K-NEAREST NEIGHBOR DAN NEIGHBOR WEIGHTED K-NEAREST NEIGHBOR UNTUK DETEKSI PENYAKIT STROKE*. Diss. University of Technology Yogyakarta, 2020.
- [3] Pradeepa, S., Manjula, K. R., Vimal, S., Khan, M. S., Chilamkurti, N., & Luhach, A. K. (2023). DRFS: detecting risk factor of stroke disease from social media using machine learning techniques. *Neural Processing Letters*, 1-19.
- [4] Singh, M. Sheetal, Prakash Choudhary, and Khelchandra Thongam. "A comparative analysis for various stroke prediction techniques." *Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27–29, 2019, Revised Selected Papers, Part II 4*. Springer Singapore, 2020.
- [5] Bandi, V., Bhattacharyya, D., & Midhunchakkravarthy, D. (2020). Prediction of Brain Stroke Severity Using Machine Learning. *Rev. d'Intelligence Artif.*, 34(6), 753-761.
- [6] Van Os, H. J., Ramos, L. A., Hilbert, A., Van Leeuwen, M., Van Walderveen, M. A., Kruyt, N. D., ... & Mr Clean Registry Investigators. (2018). Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. *Frontiers in neurology*, 9, 784.
- [7] Dataset named 'Stroke Prediction Dataset' from Kaggle: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>.
- [8] Ma'rifah, H., Wibawa, A. P., & Akbar, M. I. (2020). Klasifikasi artikel ilmiah dengan berbagai skenario preprocessing. *Ekonomi Bisnis*, 29, 23-01.
- [9] Analytics Vidhya. 2020. *Categorical Encoding One Hot Encoding vs Label Encoding*. Available at: <https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/>.

- [10] Rizal, A. A., & Soraya, S. (2018). Multi time steps prediction dengan recurrent neural network long short term memory. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 18(1), 115-124.
- [11] Documentation for Logistic Regression from Scikit-learn.org.
- [12] Grigoryev, S. G., Lobzin, Y. V., & Skripchenko, N. V. (2016). The role and place of logistic regression and ROC analysis in solving medical diagnostic task. *Journal Infectology*, 8(4), 36-45.
- [13] Documentation for Decision Tree Classification from Scikit-learn.org.
- [14] Wu, Y., Xia, Z., Feng, Z., Huang, M., Liu, H., & Zhang, Y. (2024). Forecasting Heart Disease Risk with a Stacking-Based Ensemble Machine Learning Method. *Electronics*, 13(20), 3996.
- [15] Documentation for Random Forest Classification from Scikit-learn.org.
- [16] Shimizu, G. Y., Schrempf, M., Romão, E. A., Jauk, S., Kramer, D., Rainer, P. P., ... & de Azevedo-Marques, P. M. (2024). Machine learning-based risk prediction for major adverse cardiovascular events in a Brazilian hospital: Development, external validation, and interpretability. *PloS one*, 19(10), e0311719.
- [17] Documentation for K-Nearest Neighbor from Scikit-learn.org.
- [18] Arshad, H. (2024). The Wine Quality Prediction Using Machine Learning. *Journal of Innovative Computing and Emerging Technologies*, 4(2). Documentation for Support Vector Machine from Scikit-learn.org.
- [19] Documentation for Support Vector Machine from Scikit-learn.org.
- [20] Feng, Qiya. "Investigation Based on Machine Learning Algorithms." 342.
- [21] Documentation for Naïve Bayes Classification Algorithm from Scikit-learn.org.
- [22] SALIHU, S. A., OWOYEMI, O. A., & SALIU, K. B. (2024). Performance Analysis of Some Machine Learning Algorithms in Prediction of Heart Disease. In *Conference Organising Committee* (p. 169).

# Evaluation of the Efficiency of User Interfaces for AI Tools

Gordana Jotanovic<sup>1</sup>, Goran Jausevac<sup>1</sup>, Vladimir Brtko<sup>2</sup>, Armin Podanovic<sup>1</sup>, Jelena Simic<sup>1</sup> and Ajla Nurkanovic<sup>1</sup>

<sup>1</sup> University of East Sarajevo, Faculty of Transport and Traffic Engineering, Vojvode Misica 52, Doboj, Bosnia and Herzegovina

<sup>2</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin", Djure Djakovica bb, 23000 Zrenjanin, Republic of Serbia

gordana.jotanovic@sf.ues.rs.ba; goran.jausevac@sf.ues.rs.ba; vladimir.brtka@uns.ac.rs; arminlukavac89@gmail.com; simicjelena339@gmail.com; ajlanurkanovic11@gmail.com

## Abstract:

The introduction of user interfaces for AI tools is bringing about revolutionary changes in the way users interact with technology. The efficiency of these interfaces plays a key role in the user experience, enabling simpler, more intuitive and more productive interactions. The results of the research presented in the paper clearly show that Google Gemini stands out in efficiency among the analyzed AI tools, thanks to its balanced performance in all key criteria. Despite the longest response time, its fast startup time and significant response length allow it to be the most successful. On the other hand, GPT Chat AI offers quick response and solid response length, but it is not as balanced as Google Gemini. Microsoft Copilot, despite having the fastest startup time and shortest response time, falls behind due to significantly shorter response length, which can be a drawback for users who prefer more detailed information. The study also highlights the importance of analyzing specific user needs when choosing an AI tool, as each tool offers unique advantages in different aspects.

## Keywords:

User Interface (UI), Artificial Intelligence (AI), Meta AI, Microsoft Copilot, GPT Chat AI, Google AI, Together AI, User Experience (UX), Simple Additive Weighting (SAW), Analytical Hierarchy Process (AHP).

## 1. Introduction

A key area of design is the user interface, which shapes the way users interact with technology, software applications, and digital platforms. Whether it is mobile apps, websites, or software systems, the interaction between humans and computers represents the User Interface (UI) or the space where communication between humans and machines occurs. The goal of this interaction is to enable efficient operation and control of the machine by humans, while the machine simultaneously returns information that helps the operator in the decision-making process. Design considerations applied in creating a user interface encompass disciplines such as ergonomics and psychology. Generally, the goal of user interface design is to produce a user interface that makes it easy, efficient, and enjoyable (usable) to operate a machine in a way that yields the desired result (i.e., maximum usability). This means that the operator must ensure minimal input to achieve the desired output and that the machine minimizes unwanted outputs for the user.

User interfaces consist of one or more layers, including the Human-Machine Interface (HMI), which typically connects machines with physical input hardware (such as keyboards, mice, or gamepads) and output hardware (such as computer monitors, speakers, and printers). A device that implements HMI is called a Human Interface Device (HID). User interfaces that do not use physical body movements as an intermediary step between the brain and the machine do not use input or output devices other than electrodes; these are called brain-computer interfaces (BCI) or Brain-Machine Interfaces (BMI). Composite User Interfaces (CUI) are user interfaces that interact with two or more senses. The most common CUI is the Graphical User Interface (GUI), which consists of a tactile user interface and a visual user interface that can display graphics. When sound is added to the GUI, it becomes a

multimedia user interface (MUI). There are three broad categories of CUI: standard, virtual, and augmented. Standard CUI uses standard human interface devices such as keyboards, mice, and computer monitors.

AI tools have become indispensable in many industries, providing solutions for data analysis, process automation, decision making and improving the user experience. From medical applications that aid in diagnosis to tools for predictive analytics in finance, AI tools are helping to take efficiency and precision to the next level. As AI technology advances, its application is expected to expand to even more areas, providing innovative solutions to global challenges.

The most famous AI tools are:

Meta AI is part of Meta Platforms (formerly Facebook), which focuses on developing AI and increasing user interaction on popular social platforms like Facebook, Instagram and WhatsApp [1]. Meta AI uses the open source LLaMA 3.2 model, which allows users to create, interact with, and consume AI-generated content [2]. Meta AI is available in several countries and allows users to customize content according to their needs [3].

Microsoft Copilot is an AI assistant that integrates into various Microsoft products such as Word, Excel, PowerPoint and Outlook. Copilot helps users to quickly and efficiently perform tasks, such as writing, organization, and creativity. Copilot can analyze and synthesize information from different sources, providing users with detailed and relevant answers [4].

GPT Chat AI is an AI model developed by OpenAI, which uses the GPT-4 architecture to generate text and chat [2]. GPT Chat AI can understand and respond to various user queries, providing conversational interactions that approximate human [2]. This AI model is known for its ability to generate realistic and contextually relevant responses.

Google AI is a platform that uses advanced AI technologies to improve various aspects of user experience [5]. Some of the key tools include: Magic Editor and Magic Eraser in Google Photos: These tools allow users to create and edit images using AI [6]. AI-powered dubbing in YouTube: This tool enables the translation of video content into different languages [5]. AI Search: Generative AI improves search by enabling users to get relevant answers to queries [5].

Together AI is a platform that provides tools for rapid learning, fine-tuning and implementation of AI models [7]. Some of the key tools include: Together Inference Engine: The fastest inference engine, which is 4 times faster than vLLM [8]. Together Fine-Tuning: Enables fine-tuning of generative AI models using own data [7]. Together GPU Clusters: Platform for scaling and performance optimization of AI models [7].

The paper is structured as follows: the first chapter contains an introduction to the issue. The second chapter deals with thematic considerations of studies that analyze different aspects of user experience (UX) in the context of AI tools, including user satisfaction. The third chapter is about testing the effectiveness of the user interface for AI tools. The research results are described in the fourth chapter. Concluding considerations and future research development are presented in the fifth chapter.

## 2. Previous work

The application of AI technologies in everyday life and work, it becomes imperative that user interfaces are intuitive and efficient. The introduction to this topic includes an overview of previous studies that analyze various aspects of user experience (UX) in the context of AI tools, including user satisfaction, ease of use, and functionality. Studies focus on design methods, evaluation metrics, and experimental approaches used to evaluate the performance of user interfaces. Understanding these factors is essential to developing AI systems that are not only powerful, but also easily accessible and useful to end users.

The study [9] presents evaluator performance and the observation that evaluators under similar conditions identify significantly different sets of usability problems. The study consists of 50 usability tests by novice evaluators and subsequent comparisons, in teams and individually, of the resulting usability problems. The same problems were independently analyzed by 10 human-computer interaction experts. The study shows an agreement between evaluators of about 40%, which indicates a significant evaluator effect.



Study [10] contributes to a deeper understanding of user experiences and offers valuable guidance for improving the usability of generative AI applications. The research presents a usability evaluation and comparison of generative AI applications through the analysis of user reviews from popular digital markets, especially the Apple App Store and Google Play. The study aims to bridge the research gap in real-world usability evaluations of generative AI tools. A total of 11,549 reviews were extracted and analyzed from January to March 2024 for five generative AI applications: ChatGPT, Bing AI, Microsoft Copilot, Gemini AI and Da Vinci AI. The results show that ChatGPT achieved the highest composite usability scores among Android and iOS users, with scores of 0.504 and 0.462, respectively. On the other hand, Gemini AI had the lowest score among Android apps at 0.016 and Da Vinci AI had the lowest score among iOS apps at 0.275.

The study [11] was tasked with finding, evaluating and synthesizing existing state-of-the-art work in the field of Intelligent User Interfaces (IUI) with an additional focus on IUI evaluation. The research analyzed 211 studies published in this field between 2012 and 2022. The studies are mostly related to Human-Computer Interaction (HCI) and Software Engineering (SE) domains. Evaluation of IUI is mainly carried out through experiments and questionnaires, although usability and User Experience (UX) are not considered together in evaluations. The majority of evaluations (81% of studies) reported a partial or full improvement in usability or user experience.

The goal of the research is to map the use of artificial intelligence in the UX design process. New digital tools such as Artificial Intelligence (AI) incorporated into the UX process have the potential to improve efficiency and accuracy, while creating more innovative and creative solutions. The result of the research was that understanding how artificial intelligence can be used for UX has important research and practical implications [12].

The study provides guidelines for human-AI interaction, with recommendations for effectively rejecting the user, modifying or otherwise effectively recovering from AI choices that the user does not like. According to [13], we need AI-resilient interfaces that help people be resilient to AI choices that aren't right for them from a user perspective.

The paper [14] explores how to properly design user interface modules in order to achieve user satisfaction of a mobile application. Introducing FEELER, a framework for fast and intelligent exploration of user interface module design solutions with a collective machine learning approach. The task of FEELER is to help designers to quantitatively measure the preference score of different design solutions, with the aim of facilitating designers to adapt user interface modules appropriately and quickly. The research was conducted on real cases of user interface module design in the Baidu application, which is one of the most popular mobile applications in China.

In the study, an automated data warehousing method was developed to generate suitable data for UI-Hawk training. The experiment was done on FunUI and the GUI navigation benchmarks consistently confirm that understanding screen flow is not alone useful, but also essential for GUI navigation [15].

In this paper [16], a systematic literature review was carried out to investigate IUI design trends in the context of modern software systems, such as software systems based on the Internet of Things (IoT) or dedicated to smart cities. The research results show which models and technologies are most used for the development of IUI and which application domain is the most represented.

### 3. Testing the efficiency of the User Interface for AI tools

Evaluating the effectiveness of the user interface for AI tools is extremely important to ensure that users have a positive and productive experience when interacting with the technology. The test includes different methods and metrics to evaluate different aspects of the user experience. The research presented in the paper deals with the examination of User Experience (UX): The goal of the research is to choose the most effective AI tool from the offered tools in relation to UI. It was analyzed how users perceive the interface, how easy it is for them to use, how quickly they can perform tasks and how long the system's responses to set tasks are.

The UX test was performed using the SAW (Simple Additive Weighting) multi-criteria decision-making method using several steps, see Figure 1.

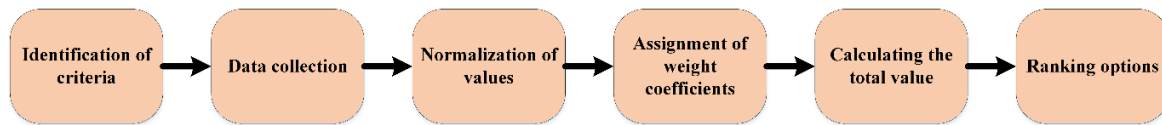


Figure 1. Steps of testing user interface efficiency for AI tools

- **Identification of criteria:** The first step is the identification of criteria that will be used to evaluate the efficiency of the interface, such as ease of use, speed of performing tasks, aesthetic design, adaptability to users with disabilities, and user satisfaction.
- **Data collection:** Collection of relevant data through surveys, user tests, and interface performance analysis.
- **Normalization of values:** The values of the collected data are normalized in order to enable their comparability. This includes scaling the results to a common scale.
- **Assignment of weight coefficients:** Each criterion is assigned weight coefficients based on their importance. Weight coefficients can be determined by expert opinion or user preferences.
- **Calculation of the total value:** The normalized values for each option are multiplied by the corresponding weighting coefficients and then added to obtain the total value for each option.
- **Ranking of options:** Based on the total values, the options are ranked from the most efficient to the least efficient user interface.
- This approach enables comprehensive analysis and identification of best practices in user interface design for AI tools, thereby improving user experience and increasing efficiency.

## 4. Results and discussion

SAW and AHP methods were used to test the effectiveness of user interfaces for AI tools. The methods allow detailed evaluation of various aspects of the interface. Also, SAW and AHP methods are used to evaluate options based on defined criteria, enabling systematic analysis and performance comparison. AI tools were tested by measuring user experience in the form of the time it takes to load the user interface for the first time, the time it takes for the AI tool to respond to a user request, and the length of the AI tool's response to the same question "How long does it take to test the response speed of the AI tool?". The task of user interface testing is to choose the most effective AI tool. When choosing the most effective AI tool in relation to UI, three alternatives were defined in the research: Google Gemini, Microsoft Copilot and GPT Chat AI. Also, criteria are defined: user interface startup time, response length and response time, see Table 1.

**Table 1:**

Values of the criteria in relation to the set alternatives

	UI startup time (s)	Length of response (Word Count)	Response time (s)
Google Gemini	1,57	294	11,45
Microsoft Copilot	1,31	56	03,76
GPT Chat AI	1,50	132	15,17

**UI startup time** is the time it takes for the application, if the application is opened for the first time, to display the UI, measured in seconds (s). Gemini has a UI startup time of 1.57 (36%), Microsoft Copilot of 1.31 (34%) and GPT Chat AI of 1.50 (30%).

**Length of response** represents the length of the response in the form of a text message that we received to the question "Koliko iznosi vrijeme potrebno za testiranje brzine odziva AI alata?". The same question was asked for consideration by all three AI tools and the number of words answered to the question was measured (Word Count). **Length of response** values are: Gemini (294; 61%); Microsoft Copilot (56; 12%) and GPT Chat AI (132; 27%).

**Response time** indicates the time required for the application to respond to the user's request. In the context of AI tools, it is the time that elapses from the moment the user enters a query or command to the moment he receives an answer or result. A short **Response time** is crucial for a good user interface because it directly affects the efficiency and satisfaction of users. The efficiency values are: Gemini (11.45; 38%); Microsoft Copilot (3.76; 12%) and GPT Chat AI (15.17; 50%).

The criteria for the SAW method were weighted. However, different weightings of the criteria can lead to different results. The weighted values of the criteria we adopted are: **UI startup time** 0.3; **Length of response**: 0.4; **Response time**: 0.3. We normalize ratings by minimizing **UI startup time** and **Length of response**, according to (1).

$$rij = \min_{i,j} \{a_{ij}\} / a_{ij} \quad rij_{ij} = \frac{\min(a_{ij})}{a_{ij}} \quad rij = a_{ij} \min(a_{ij}) \quad (1)$$

Where is:

- $rij_{ij}$  normalized grade for alternative iii according to criterion jjj
- $a_{ij}$  original evaluation of alternative iii according to criterion jjj

We maximize the normalization of ratings for **Response time**, according to (2).

$$rij = a_{ij} \max_{i,j} \{a_{ij}\} / a_{ij} \quad rij_{ij} = \frac{a_{ij}}{\max(a_{ij})} \quad rij = \max(a_{ij}) a_{ij} \quad (2)$$

Where is:

- $rij_{ij}$  normalized grade for alternative iii according to criterion jjj
- $a_{ij}$  original evaluation of alternative iii according to criterion jjj

Values were normalized, see Table 2.

**Table 2:**

Normalized values for SAW method

	UI startup time	Length of response	Response time
Google Gemini	≈0.637	1	≈0.690
Microsoft Copilot	≈0.763	≈0.487	≈0.266
GPT Chat AI	≈0.667	≈0.401	≈0.193

For each alternative, a weighted sum of normalized grades is calculated, according to (3):

$$Si = \sum_{j=1}^n w_j \cdot rij_{ij} \quad Si = \sum_{j=1}^n w_j \cdot rij \quad (3)$$

Where is:

- $Si$  overall rating of alternative iii
- $w_j$  weight of criteria jjj
- $rij_{ij}$  normalized assessment of alternative iii according to criterion jjj

Alternatives are ranked on the basis of overall  $Si$  scores, see Table 3.

**Table 3:**

Calculation of the weighted sum for SAW method

	UI startup time	Length of response	Response time	$Si$ overall score
Google Gemini	0,191	0,4	0,026	0,617
Microsoft Copilot	0,229	0,076	0,080	0,385
GPT Chat AI	0,200	0,180	0,058	0,438

The ranking of alternatives is done on the basis of the overall score  $Si$  for SAW method.  $Si$  scores for specific alternatives are: Google Gemini (0,617; 43%); Microsoft Copilot (0,385; 27%) and GPT Chat AI (0,438; 43%), Figure 2.

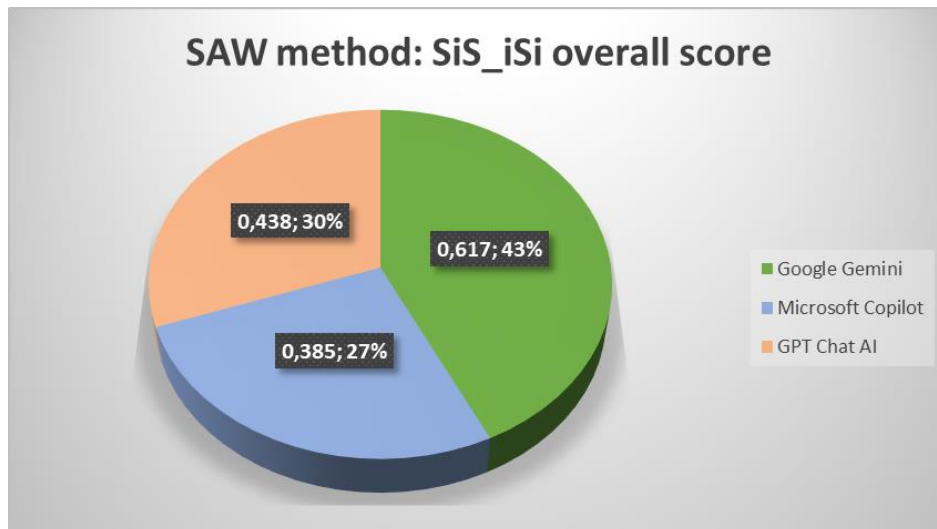


Figure 2. SAW method: Rating of SiS\_iSi

The results of the SAW method give us insights into the effectiveness of different AI tools based on the given criteria. With the highest overall score of 0.617, Google Gemini shows a balanced performance across all three criteria. Although it has the longest response time, its fast **UI startup time** and large **Length of response** allow it to dominate the overall rating. Based on an overall score of 0.438, GPT Chat AI offers **Response time** and a solid **Length of response**. Its **UI startup time** is also quite fast, making it a reliable option, but it is not as balanced as Google Gemini. Microsoft Copilot with a score of 0.385, has the fastest **UI startup time** and the shortest **Response time**, but the **Length of response** is significantly shorter compared to other tools. This can be a handicap if users prefer more detailed answers.

**Google Gemini** stands out as the most effective tool for the given criteria, while GPT Chat AI and Microsoft Copilot offer specific advantages in different aspects.

For more accurate analysis, in addition to the SAW method, the Analytical Hierarchy Process (AHP) method of multi-criteria decision-making was performed with the same data [17, 18]. The same criteria as for the SAW method have been established for the AHP method. Then weights were assigned based on the importance of each criterion: **UI startup time**: 0.2; **Length of response**: 0.4; and **Response time**: 0.4. The values were normalized and the weighted sum was calculated, see Table 4.

**Table 4:**

Calculation of the weighted sum for AHP method

	UI startup time	Length of response	Response time	SiS_iSi overall score
Google Gemini	0.127	0.400	0.035	0.562
Microsoft Copilot	0.153	0.076	0.106	0.335
GPT Chat AI	0.133	0.179	0.077	0.389

The ranking of alternatives is done on the basis of the overall score SiS\_iSi. SiS\_iSi for AHP method scores for specific alternatives are: Google Gemini (0.562; 44%); Microsoft Copilot (0.335; 26%) and GPT Chat AI (0.389; 30%), Figure 3.

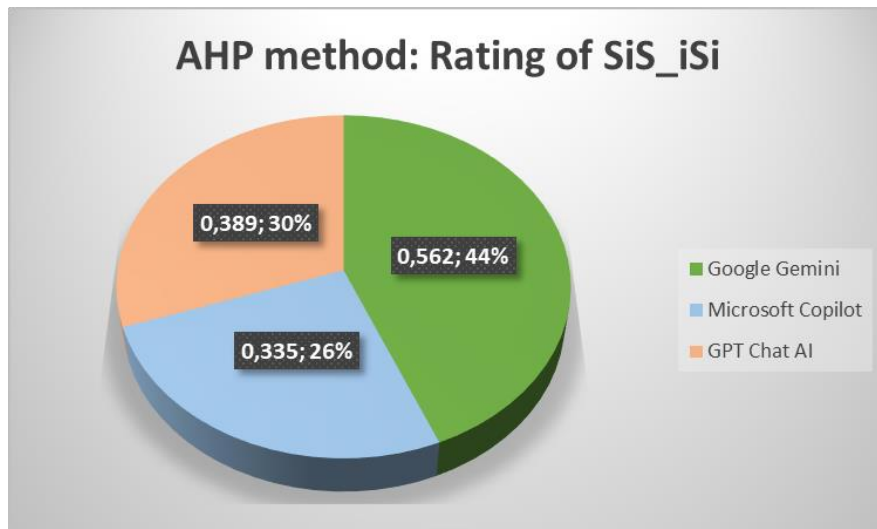


Figure 3. AHP method: Rating of SiS\_iSi

Comparative Analysis of Google Gemini: SAW score of 0.617, AHP score of 0.562. Consistently strong performance across criteria. GPT Chat AI: SAW score of 0.438, AHP score of 0.389. Solid but not as balanced as Gemini. Microsoft Copilot: SAW score of 0.385, AHP score of 0.335. Fast but lacks detail in responses.

Both metrics agree that **Google Gemini** leads in overall efficiency, while GPT Chat AI and Microsoft Copilot have specific strengths but lag behind in balance and detail. These comparisons validate the robustness of **Google Gemini** in various evaluation frameworks.

## 5. Conclusions

The results of the SAW and AHP methods clearly show that Google Gemini leads in efficiency among the analyzed AI tools, thanks to balanced performance in all key criteria. Although it has the longest response time, its quick start-up and long response time allow it to be the most successful. On the other hand, GPT Chat AI offers quick response and solid response length, but is not as balanced as Google Gemini. Microsoft Copilot, despite the fastest start-up and the shortest response time, lags behind due to a significantly shorter response length, which can be a disadvantage for users who prefer more detailed information. Also, the research emphasizes the importance of analyzing the specific needs of users when choosing an AI tool, because each tool offers unique advantages in different aspects. A shortcoming of the research is that certain versions of the AI tool were not taken into account when measuring, which would probably result in different values.

Future research in the area of evaluating the effectiveness of user interfaces for AI tools could lead to the development of new metrics and methods for measuring the effectiveness of user interfaces, including time metrics, user satisfaction, and task performance. Effectiveness testing with a larger number of questions for AI tools should also be taken into account in further research.

Research could go in the direction of developing adaptive UI design, which can dynamically adjust its elements based on the user's behavior and preferences. The development of interactions through natural language allows users to communicate with AI tools through speech and text, and how these tools can better understand and respond to user queries. Improvement of multimodal interfaces that represent a combination of different sensory inputs (eg voice, touch, gestures) can serve to create more intuitive and accessible interfaces.

In conclusion, the research presented in the paper should contribute to improving the efficiency of AI tools in relation to AI. Which enables AI tools to be more efficient, intuitive and user-friendly.

## References:

- [1] Trend Micro Incorporated, "Trend Micro Blog," Trend Micro News. Accessed: Oct. 31, 2024. [Online]. Available: <https://news.trendmicro.com/>
- [2] Future US, Inc., "Tom's Guide | Tech Product Reviews, Top Picks and How To," Tom's Guide. Accessed: Oct. 31, 2024. [Online]. Available: <https://www.tomsguide.com>
- [3] Meta, "Meta | Social Metaverse Company," Meta | Social Metaverse Company. Accessed: Oct. 31, 2024. [Online]. Available: <https://about.meta.com/>
- [4] Microsoft, "Microsoft – AI, oblak, produktivnost, računarstvo, igranje igara i aplikacije," Microsoft. Accessed: Oct. 31, 2024. [Online]. Available: <https://www.microsoft.com/bs-latn-ba>
- [5] J. Nielsen, *Usability engineering*. Morgan Kaufmann, 1994.
- [6] D. Norman, *The design of everyday things: Revised and expanded edition*. Basic books, 2013.
- [7] E. R. Tufte, "The visual display of quantitative information," *The Journal for Healthcare Quality (JHQ)*, vol. 7, no. 3, p. 15, 1985.
- [8] B. Shneiderman, C. Plaisant, M. S. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos, "Designing the User Interface: Strategies for Effective Human-Computer Interaction," 2016.
- [9] K. Hornbæk and E. Frøkjær, "A study of the evaluator effect in usability testing," *Human-Computer Interaction*, vol. 23, no. 3, pp. 251–277, 2008.
- [10] R. Alabduljabbar, "User-centric AI: evaluating the usability of generative AI applications through user reviews on app stores," *PeerJ Computer Science*, vol. 10, p. e2421, 2024.
- [11] S. Brdnik, T. Heričko, and B. Šumak, "Intelligent user interfaces and their evaluation: a systematic mapping study," *Sensors*, vol. 22, no. 15, p. 5830, 2022.
- [12] Å. Stige, E. D. Zamani, P. Mikalef, and Y. Zhu, "Artificial intelligence (AI) for user experience (UX) design: a systematic literature review and future research agenda," *Information Technology & People*, 2023.
- [13] E. L. Glassman, Z. Gu, and J. K. Kummerfeld, "AI-Resilient Interfaces," *arXiv preprint arXiv:2405.08447*, 2024.
- [14] J. Zhou *et al.*, "Intelligent exploration for user interface modules of mobile app with collective learning," presented at the Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3346–3355.
- [15] J. Zhang, Y. Yu, M. Liao, W. Li, J. Wu, and Z. Wei, "UI-Hawk: Unleashing the Screen Stream Understanding for GUI Agents," *Preprints*, Aug. 2024, doi: 10.20944/preprints202408.2137.v1.
- [16] T. G. Gonçalves, C. Kolski, K. M. de Oliveira, G. H. Travassos, and E. G.-L. Strugeon, "A systematic literature review on intelligent user interfaces: preliminary results," presented at the Adjunct Proceedings of the 31st Conference on l'Interaction Homme-Machine, 2019, pp. 1–8.
- [17] M. Wahyuni and H. Hidayatullah, "Pemilihan Toko Laptop Terbaik Di Aek Kanopan dengan Sistem Pendukung Keputusan menggunakan Metode Simple Additive Weighting," *Jurnal Sistem dan Teknologi Informasi Cendekia (JuSTICE)*, vol. 1, no. 2, pp. 54–60, 2024.
- [18] Z. H. B. Huang, "Evaluating MCDM Methodologies for Data Integration Tool Selection: A Comparative Study of COPRAS and NR-TOPSIS," Engineering, The George Washington University, 2024. Accessed: Apr. 11, 2024. [Online]. Available: <https://www.proquest.com/openview/6c6b09d932cf004f5dcda447350b10bb/1?pq-origsite=gscholar&cbl=18750&diss=y>

# Integration of Image Recognition Systems in Web Applications

Nikola Jovanov<sup>1</sup>, Eleonora Brtko<sup>1</sup>, Vesna Makitan<sup>1</sup>, Predrag Novokmet<sup>1</sup> and Albina Kószó<sup>2</sup>

<sup>1</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin" Djure Djakovica bb, Zrenjanin, Republic of Serbia

<sup>2</sup> University of Szeged, Dugonics tér 13, Szeged, Hungary

nikola.jovanov@tfzr.rs; eleonora.brtka@tfzr.rs; vesna.makitan@tfzr.rs; predrag.novokmet@tfzr.rs; albina.koszo@gmail.com

## Abstract:

This paper presents the integration of image recognition systems in web applications, focusing specifically on the ease and speed of implementing the MobileNet model using TensorFlow.js within a React application. The primary objective is to demonstrate how straightforward it is to incorporate advanced image recognition capabilities into modern web development. The methodology includes a brief theoretical exploration of TensorFlow.js and React, followed by the development of a practical application that validates the proposed approach. The results highlight the effectiveness and efficiency of the MobileNet model, showcasing quick loading times and accurate image recognition of simple images showing a singular object, while recommending the use of object detection when it comes to more complex images.

## Keywords:

Image Recognition, TensorFlow, React, MobileNet, Trained Machine Learning Model

## 1. Introduction

Machine learning has advanced significantly over the past few decades [1], bringing various functionalities to the forefront of technology. One such functionality is image recognition, which serves as the primary focus of this study. Image recognition is prevalent across numerous domains, including medicine [2], traffic [3], and security [4]. This research utilizes a universal model that has been trained on a vast array of diverse data, showcasing high performance when it comes to execution speed, as well as ease of implementation [5]. The effectiveness of this model is primarily tested through image recognition of animals.

This study encompasses both theoretical and empirical research. The theoretical component explains technologies used in this work's solution development, while the empirical research involves the development of the application and an analysis of the development process and the results obtained. The integration of the MobileNet model for image recognition within the React web application is detailed, followed by an evaluation of the model's accuracy against the input images, as well as explanation of its ease of use.

The aim of this work is to illustrate how this integration is achieved and how it facilitates and accelerates the use of complex systems such as machine learning models for image recognition. The hypothesis of this research posits that "the use of pre-trained machine learning models simplifies and accelerates the development of web applications for image recognition".

## 2. Theoretical research

Theoretical research includes basics of TensorFlow platform and React JavaScript library. React works as a basis for enabling development of an application that will integrate pre-trained TensorFlow model named MobileNet, as well as provide environment for invocation of its function for image recognition.

Topics explained in theoretical research will be implemented in application development.

## 2.1. TensorFlow

TensorFlow is an open-source machine learning library developed by Google in 2015 [6]. Since its launch, it has become one of the most popular platforms for developing and implementing machine learning models [7]. Known for its versatility, TensorFlow supports a range of tasks, from simple regression models to complex deep neural networks [8].

This library is used in various fields, including natural language processing (NLP), image recognition, and recommendation systems. Its flexibility and scalability enable researchers and engineers to quickly develop and test different models, while support for both, central processing units (CPU) and graphics processing units (GPU), as well as tensor processing units (TPU), accelerates training via large datasets [9].

Key features that make TensorFlow suitable for machine learning development include:

- Flexibility – Users can choose between different levels of abstraction, utilizing high-level APIs like Keras for simpler tasks or working directly with low-level APIs for greater control [9].
- Cross-platform support – TensorFlow can be utilized across various platforms, including Windows, macOS, and Linux, as well as on mobile devices and in web applications [10].
- Distributed learning – TensorFlow supports distributed model training across multiple machines, which is advantageous for large projects and working with extensive datasets [11].

Some of the key features of TensorFlow include:

- Tensors – basic units of data in TensorFlow. They represent multidimensional arrays (similar to matrices) and are used to represent data in various forms. Tensors can be scalars, vectors, matrices, or multi-dimensional arrays [12].
- Graphs – structure for building and executing machine learning models. In TensorFlow, models are represented as directed graphs, where nodes are operations (such as multiplication, addition, activation), and tensors are the units transmitted between operations. This graphical approach allows for efficient execution and optimization of computations [13].
- Models – represent mathematical functions that map input data to their output predictions. The model used in this project is a pre-trained model, as TensorFlow provides the option to use already trained models that have been trained on large datasets. This allows for faster application development and greater accuracy in data recognition [14].

This work utilizes TensorFlow.js, which is a JavaScript library that allows for the execution and training of machine learning models directly in the browser or on a Node.js server. This library enables developers to leverage machine learning in web applications without the need for complex infrastructure. It provides access to a wide range of tools for creating, training, and executing models, allowing for the implementation of high-quality functionalities directly in frontend applications [15].

## 2.2. React

React.js is a JavaScript library developed by Meta (formerly Facebook) used for building user interfaces for web applications. Released in 2013, the React library quickly gained popularity due to its efficiency and flexibility [16]. The primary goal of React is to enable developers to create complex, interactive, and dynamic applications using a component-based approach. This component-based method facilitates code reusability, allowing developers to focus on individual parts of the application, making development faster and more organized [17].

JSX, or JavaScript XML, is a syntax used in React for describing the appearance of the user interface. It allows writing HTML within JavaScript code, creating a more intuitive and readable code structure. JSX is converted into regular JavaScript before execution, meaning that each JSX element is essentially a JavaScript object. This enables dynamic content generation and easy manipulation [18].

In React, everything is based on components, which are self-contained, reusable parts of user interface (UI). There are two main types of components: functional components and class components. Functional components are simple and often utilize hooks, and are considered preferred way of typing React code [18, 19].



Hooks are functions that enable the use of state and other React features within functional components [19]. The most significant hooks used in this work are `useState` and `useEffect`.

State in React is used to manage dynamic data within a component. When the state changes, React automatically re-renders the component, ensuring that the UI is always in sync with the data [18].

On the other hand, `useEffect`, is used for managing side effects, such as Application Programming Interface (API) calls or Document Object Model (DOM) manipulation. This hook runs after every render but can be configured to run only when specific properties or states change [20]. In this work, `useEffect` is used to load the image recognition model during the initial rendering of the application.

For this specific project, React is used for its dynamic rendering properties, as it fits perfectly with gradual addition of functionalities inside of the application page.

### 3. Application development

#### 3.1. User manual – summary of application functionality

Upon launching the application, users must wait for the MobileNet model to load. During this time, the main screen is not displayed. Instead, a title indicating that the model is loading is shown. This is achieved through a simple state check using the `useState` hook in React, which maintains this loading view until the model is fully loaded.

Once the loading is complete, the application's main page appears. This page is the sole interface of the application which dynamically changes based on whether an image has been uploaded or identification has been initiated.

By clicking a button, a file selection window opens, allowing users to choose files from their local computer. After selecting an image, it is displayed in a designated area, and a button which initiates the identification process appears. The name of the selected image is also shown below the selection button.

The identification button communicates with the image recognition model, resulting in a new section appearing on the screen that displays the recognition results, completing the rendering process of the web application page, as seen in the Figure 1. This section shows the top three results with the highest accuracy, with the first result (marked with the highest percentage) highlighted in a specific way.

While this concludes the application's functionality, users can select a new image from this state, which removes the results section and prompts a new identification process.

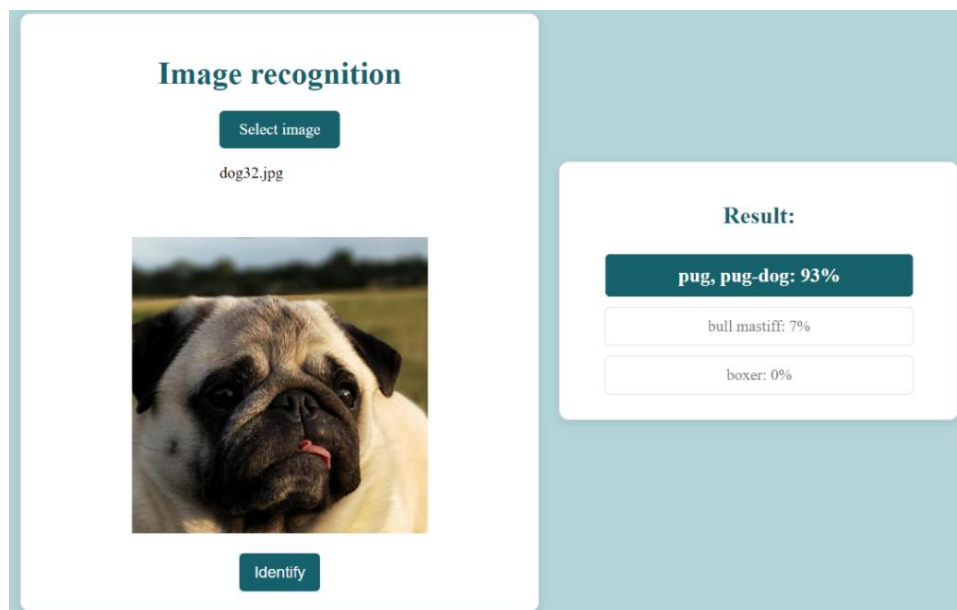


Figure 1: Final state of application

### 3.2. Uploading an image

A specific component, whose call can be seen in Listing 1, is responsible for uploading images and is called within the main application component.

```
<UploadComponent onFileChange={uploadImage} fileName={fileName} />
```

Listing 1: Upload Component call in main component

The function *uploadImage*, along with the file name, is passed as the *onFileChange* value. This function is triggered via the *onChange* event of the input tag, activating each time the selected file changes. The input type must be set to *file*, which allows uploading files for classification.

Additionally, this component includes a label with the text "Select image", which, through a modified style, displays the content as a button instead of the standard file upload field, and can be seen in the Listing 2.

```
import React from "react";
const UploadComponent = ({ onFileChange, fileName }) => {
 return (
 <div className="inputHolder">
 <label className="uploadLabel">
 Select image
 <input type="file"
 accept="image/*"
 className="uploadInput"
 onChange={onFileChange}
 />
 </label>
 {fileName && <p className="fileName">{fileName}</p>}
 </div>
);
};
```

```
export default UploadComponent;
```

Listing 2: Upload Component

At the end of the rendering, there is a check to see if the *fileName* has a value, i.e., whether it is a "truthy." This serves as a quick check to print a paragraph displaying the name of the selected file.

The *uploadImage* function resides in the main application component and, as mentioned, is passed to the image upload component. This function contains all the necessary logic for file selection, state variable changes and clearing existing predictions, and is displayed in the Listing 3.

```
const uploadImage = (e) => {
 const { files } = e.target;
 if (files.length > 0) {
 const url = URL.createObjectURL(files[0]);
 setImageURL(url);
 setPredictions([]);
 setFileName(files[0].name);
 } else {
 setImageURL(null);
 setFileName("");
 }
};
```

Listing 3: Function for image upload

### 3.3. Displaying an image

After uploading an image, role of application is to also display it, alongside its file name. To achieve this, a component for image display is called withing the main application component, like in the Listing 4.

```
<ImageDisplay imageURL={imageURL} imageRef={imageRef} />
```

Listing 4: Image Display Component call in main component

The component receives the image URL values and a reference to the image for object manipulation. A simple check is performed to see if the image value exists, followed by rendering the image on the screen. Structure of the entire component can be seen in the Listing 5.

```
import React from "react";
const ImageDisplay = ({ imageURL, imageRef }) => {
 return <div className="imageHolder">
 {imageURL && <img src={imageURL}
 alt="upload preview"
 ref={imageRef}
 />}
 </div>;
};
export default ImageDisplay;
```

Listing 5: Image Display Component

### 3.4. Integration of MobileNet model and execution of its functions

It is necessary to install the MobileNet model and the TensorFlow.js library to enable communication with the model. After installation, both the model and the library are imported into the App.js file, where they will be utilized. Importing process is shown in Listing 6.

```
import * as mobilenet from "@tensorflow-models/mobilenet";
import "@tensorflow/tfjs";
```

Listing 6: Library import of TensorFlow.js and MobileNet model

States are defined, and methods for setting values are declared. The states used in the application for communication with the model include:

- isModelLoading - A boolean value indicating whether the model is currently loading.
- model - This state holds the MobileNet model object after it has been loaded.
- predictions - An array that stores the model's predictions after invoking its image recognition method.

States, rather than variables, are used to allow for dynamic rendering of the page. These states are shown in the Listing 7.

```
const [isModelLoading, setIsModelLoading] = useState(false);
const [model, setModel] = useState(null);
const [predictions, setPredictions] = useState([]);
```

Listing 7: Definition of states used for communication with MobileNet model

The model is loaded using a function called loadModel, which is shown in the Listing 8. This function must be asynchronous, as the mobilenet.load() function requires an unpredictable amount of execution time. While the function is executing, the isModelLoading state is set to true.

```

const loadModel = async () => {
 setIsModelLoading(true);

 try {
 const loadedModel = await mobilenet.load();
 setModel(loadedModel);
 setIsModelLoading(false);
 }
 catch (error) {
 console.log(error);
 setIsModelLoading(false);
 }
};

```

Listing 8: Load Model function

The model loading function is invoked within the `useEffect` hook, which allows for its execution during the component's rendering. In this case, the model needs to be loaded only when the application starts, specifically during the initial rendering of the `App.js` component. This is achieved by passing an empty array at the end of the `useEffect`, like in the Listing 9.

```

useEffect(() => {
 loadModel();
}, []);

```

Listing 9: Function call for loading the model inside of `useEffect` hook

This concludes the integration of `MobileNet` model, and means it is ready for use.

Once the model is loaded and the application is set up to receive data from it, a system for initiating image recognition needs to be established. The first element of this system is a button, that can be viewed in the Listing 10, that triggers the classification process. This button appears on the screen after an image is uploaded and is responsible for calling the `identify` function, seen in the Listing 11.

```

{imageUrl && (
 <button className="button" onClick={identify}>
 Identify
 </button>
)}

```

Listing 10: Button which triggers image identification

```

const identify = async () => {
 if (model && imageRef.current) {
 const predictions = await model.classify(imageRef.current);
 setPredictions(predictions);
 }
};

```

Listing 11: `Identify` function calling `classify` method of `MobileNet` model

It is necessary to provide a location to display the predictions. This is achieved through a "truthy" check, which renders the `Predictions` component only if the array exists. This check is shown in the Listing 12.

```

{predictions.length > 0 && <Predictions predictions={predictions} />}

```

Listing 12: Render call for `Predictions` Component

The role of the Predictions component, seen in the Listing 13, is to display the prediction results and ensure proper formatting of the output. When it receives an array of predictions, it uses the map function to iterate through the array, creating a new div element for each prediction.

```
import React from "react";
const Predictions = ({ predictions }) => {
 return (
 <div className="predictions">
 <h2>Result:</h2>
 {predictions.map((prediction, index) => (
 <div className={`predictionItem ${index === 0 ? "primary" : ""}`} key={index}>
 {prediction.className}: {Math.round(prediction.probability * 100)}%
 </div>
))}
 </div>
);
};
export default Predictions;
```

Listing 13: Predictions Component

In the example above (Listing 13), prediction.className is a MobileNet function that returns model's prediction, while prediction.probability returns a value representing model's certainty.

The development of the application concludes with testing the image recognition system on various types of images. For images containing a single object, the application produced satisfactory results, while images with multiple objects encountered issues in recognition, which was anticipated.

## 4. Conclusion and further work

The integration of the MobileNet model for image recognition has proven to be very simple and fast. TensorFlow.js is a highly advanced library, and it is possible to find a large number of different instructions on the internet. This has confirmed the hypothesis of the work.

The performance of the MobileNet model has also proven to be very high-quality, with the model loading in just a few seconds, and results being provided almost instantly.

The results of recognizing the content of images containing simple and clear objects were successful, while predictions from images with multiple objects yielded uncertain or incorrect results. Since images with multiple objects produced incorrect results, it is logical that the next step in the application development would be the implementation of object detection, which would segment the content of the image in a way that allows the MobileNet model to recognize multiple objects from the image and provide more accurate results.

### References:

- [1] Domingos P., "The Master Algorithm", Basic Books, 2015.
- [2] Ling-Fang Li, Xu Wang, Wei-Jian Hu, Neal N. Xiong, Yong-Xing Du, Bao-Shan Li, "Deep Learning in Skin Disease Image Recognition: A Review", IEEE Publishing, 2020.
- [3] Jingxuan Yao, "The effect of image recognition traffic prediction method under deep learning and naive Bayes algorithm on freeway traffic safety", Journal of Ambient Intelligence and Humanized Computing, 2020.
- [4] K.W. Bowyer, "Face recognition technology: security versus privacy", IEEE Publishing, 2004.
- [5] "MobileNet", Medium, URL: <https://medium.com/@nocodingai/mobilenet-fc34af9f58a5>, Accessed: October 14, 2024.
- [6] Shukla, N., "Machine Learning with TensorFlow", Manning Publications, 2018.
- [7] "TensorFlow vs. PyTorch: Which Deep Learning Framework is Right for You?", Stackify, URL: <https://stackify.com/tensorflow-vs-pytorch-which-deep-learning-framework-is-right-for-you/>, Accessed: October 14, 2024.

- [8] “Understanding TensorFlow: A Deep Dive into Machine Learning with Ease”, Medium, URL: <https://medium.com/@halfdeb/understanding-tensorflow-a-deep-dive-into-machine-learning-with-ease-4212e0a1a8f8>, Accessed: October 14, 2024.
- [9] “TensorFlow”, Databricks, URL: <https://www.databricks.com/glossary/tensorflow-guide>, Accessed: October 8, 2024.
- [10] “TensorFlow”, NVIDIA, URL: <https://www.nvidia.com/en-us/glossary/tensorflow/>, Accessed: October 8, 2024.
- [11] “Distributed training with TensorFlow”, Official TensorFlow Documentation, URL: [https://www.tensorflow.org/guide/distributed\\_training](https://www.tensorflow.org/guide/distributed_training), Accessed: October 8, 2024.
- [12] “Introduction to Tensors”, Official TensorFlow Documentation, URL: <https://www.tensorflow.org/guide/tensor>, Accessed: October 8, 2024.
- [13] “Introduction to graphs and tf.function”, Official TensorFlow Documentation, URL: [https://www.tensorflow.org/guide/intro\\_to\\_graphs](https://www.tensorflow.org/guide/intro_to_graphs), Accessed: October 8, 2024.
- [14] “Models & datasets”, Official TensorFlow Documentation, URL: <https://www.tensorflow.org/resources/models-datasets>, Accessed: October 8, 2024.
- [15] “TensorFlow JavaScript”, Official TensorFlow Documentation, URL: <https://www.tensorflow.org/js>, Accessed: October 8, 2024.
- [16] “History of ReactJS”, MyGreatLearning, URL: <https://www.mygreatlearning.com/react-js/tutorials/history-of-reactjs>, Accessed: October 14, 2024.
- [17] Álvarez-Acebal, N., “From JavaScript to React.js: Best Practices for Migration”, Internationale Hochschule Bad Honnef, 2020.
- [18] Wieruch, R. “The Road to React”, The Pragmatic Bookshelf, 2018.
- [19] “Build-in React Hooks”, Official React Documentation, URL: <https://react.dev/reference/react/hooks>, Accessed: October 14, 2024.
- [20] “useEffect”, Official React Documentation, URL: <https://react.dev/reference/react/useEffect>, Accessed: October 14, 2024.

# Cloud services modeling for long-term intellectual capital protection

Saso Nikolovski<sup>1</sup>, Bozidar Milenkovski<sup>1</sup>, Anita Petreska<sup>1</sup> and Daniela Slavkovska<sup>2</sup>

<sup>1</sup> Faculty of Information and Communication Technologies – Bitola, North Macedonia

<sup>2</sup> UTMS Skopje, Faculty of Computer Science, Republic of North Macedonia

sasnik@gmail.com; bozidar.milenkovski@uklo.edu.mk; anita.petreska1@gmail.com; dslavkovska@yahoo.com

## Abstract:

This paper presents a comprehensive performance evaluation framework to aid in the selection of optimal solutions for safeguarding and maintaining organizational data and information systems. The study emphasizes identifying critical factors and stages necessary for choosing a cloud-based solution tailored to business continuity and the protection of intellectual capital. The proposed framework facilitates selecting an adaptable recovery approach aligned with organizational operational needs, criticality of assets, and predefined timelines for business continuity and disaster recovery. Through detailed analysis and insights, this work supports informed decision-making in the implementation of reliable cloud services for resilience against outages or catastrophic events, ensuring sustained protection of an organization's intellectual capital.

## Keywords:

Intellectual capital, disaster recovery, reliability, cloud service, business continuity, data protection, data recovery.

## 1. Introduction

In today's technology-driven business environment, companies increasingly depend on a robust Business Continuity Plan (BCP) to safeguard their intellectual capital. The development of these plans aims to ensure rapid recovery procedures following an outage or disaster. The unavailability of critical systems and services can have severe repercussions, including data loss and customer dissatisfaction, which ultimately impacts revenue and the value of an organization's intellectual assets.

From the aspect of the modern digitalized work environment, organizations aspire to achieve zero downtime during operational disruptions to ensure continuity and protect their intellectual capital. However, this ideal is often unattainable due to various potential disruptions, such as weather events or cyber attacks, despite the availability of numerous recovery solutions. These solutions range from on-premises data centers to cloud-based systems[1]. Consequently, organizational management increasingly prioritizes minimizing the impact of outages by assessing the maximum acceptable downtime that the business can sustain without compromising its intellectual assets or risking long-term operational consequences.

Establishing an effective disaster recovery system that meets recovery plan requirements involves selecting solutions that align with specific utilization thresholds to uphold planned performance metrics. Prior research has often evaluated complex recovery solutions using a limited set of parameters, overlooking factors that directly influence the effectiveness of these implementations[2][3][4]. These include data transfer volume, system load during replication and recovery, and other conditions crucial for maintaining performance benchmarks.

Our research contrasts with prior studies by analyzing real-world data from an operational data center. We utilized these empirically validated parameters to develop a System Dynamics model, enabling us to assess how recovery solutions perform under various extreme conditions. This approach offers a deeper understanding of how to protect intellectual capital by ensuring reliable access to critical information assets in the face of potential disruptions.

## 2. Intellectual capital as the most important asset of the modern organizations

The development of knowledge, that is its embodiment in intellectual capital, today is a condition for economic, technological and any other form of progress in the knowledge based economy. The new way of creating values in the knowledge economy, through the management of intellectual capital and the continuous investment in it, points to the fact that for a modern organization is far more important the ability of employees to create value than the value of its tangible assets per se.

The intellectual capital of an organization represents its intangible assets as an important part of its total assets. This capital has a specific power creatively to turn the various types of knowledge, skills, structures, procedures, processes, technologies, etc. within an organization into products that have real value. The main components of intellectual capital are the human capital (managers and employees education, skills, experiences, trainings etc.), the structural capital (licenses, patents, copyrights, software, databases, organizational culture, organizational structure etc.) and the relational capital (relations with consumers, distributors, suppliers, investors, trademark, brand etc.). Each of these elements of intellectual capital contributes significantly to the success of the modern organization.

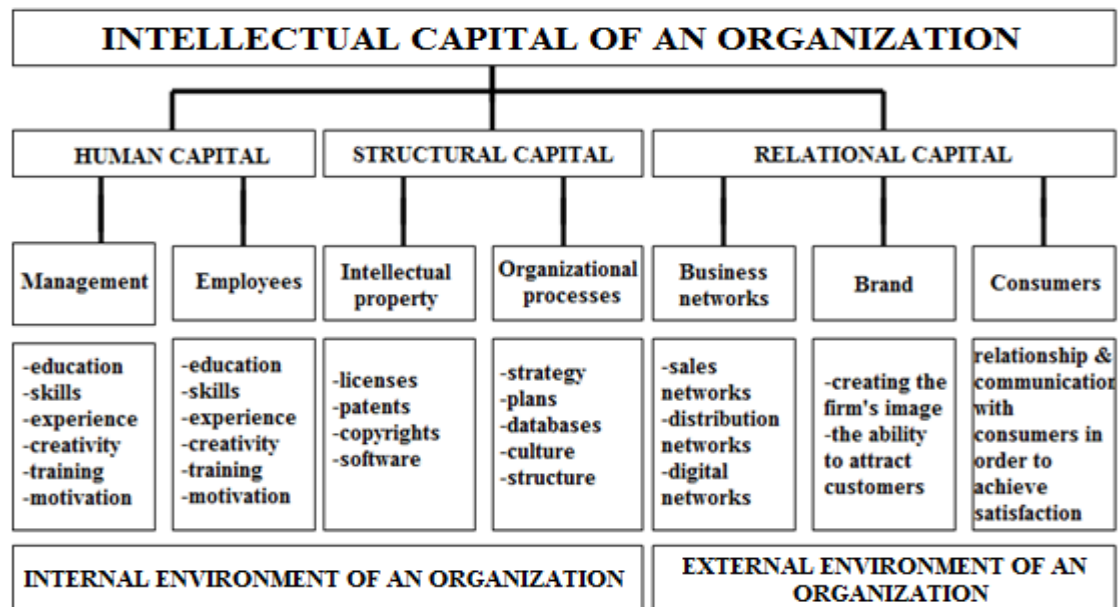


Figure 1: The intellectual capital structure of a modern organization [5]

The essence of an organizational intellectual capital lies in the value creation process. Value in an organization can be created when human capability (human capital) creates new business processes (structural capital), resulting in better products for consumers and increasing their loyalty (relational capital). Moreover, the interaction between the three constituent elements of intellectual capital is unique and unlike the usual material goods, the nature of intellectual capital is characterized by the synergistic effect. Hence, modern organizations must invest a huge effort in designing an appropriate information systems infrastructure that will be a significant motivator for creating an organization based on intellectual capital.

The importance of the intellectual capital for the modern organizations can be observed from another point of view as well. Namely, technological progress happens because organizations or individual inventors, in the desire to maximize profits, search for new and better discoveries. The opportunity to make a profit is what makes companies and entrepreneurs develop the computer, or produce a handheld camera, or produce calorie-free ice cream. Patents, copyrights, trademarks, etc., are legal mechanisms that guarantee the inventor a monopoly profit for a certain period of time. Without such mechanisms for the protection of intellectual property, it would be difficult to ensure the motivation of organizations and entrepreneurs for research work and development. So, it should be taken into consideration the fact that intellectual property, as a part of intellectual capital, is an



important determinant for the development of the modern organizations. In addition, the market value of the organization can be easily undermined if intellectual capital is not properly protected.

### 3. Previous work

The research detailed in this study builds upon a decade of work that explores the role of cloud services in supporting the daily operations of corporate entities, particularly in protecting intellectual capital [2][6][7][8][9][10][11]. A primary focus across these studies has been on key metrics like the Recovery Point Objective (RPO) and Recovery Time Objective (RTO), both of which serve as critical indicators of a data protection system's performance and reliability [12]. These parameters are especially relevant for intellectual capital protection, as they directly impact the continuity of access to vital knowledge assets during disruptions.

It is important to note that much of this research was conducted in simulated conditions, where key parameters were often derived from isolated environments, unaffected by other infrastructure elements. As such, the authors highlight that validating these results within a real-world production environment is essential to fully understanding their effectiveness in safeguarding intellectual capital.

### 4. Methodology

When evaluating outages and establishing recovery goals that align with organizational needs, it's essential to recognize that the process hinges on minimizing the time the organization is unable to operate. This process involves two key time-dependent elements. The first, a technical component, is the RTO, which defines the time required to restore systems, data, and network infrastructure. The second is Work Recovery Time (WRT), an organizationally focused measure that represents the time needed to fully reinstate operational processes. Together, these elements define the Maximum Tolerable Downtime (MTD), as outlined in the Business Continuity Plan (BCP) and Disaster Recovery Plan (DRP). The MTD is calculated as the sum of RTO and WRT:

$$MTD = RTO + WRT \quad (1)$$

As shown in Figure 2, the RTO represents the duration needed to address technical recovery, while the remaining time up to the MTD is allocated to WRT, during which all information-based work processes are fully restored.

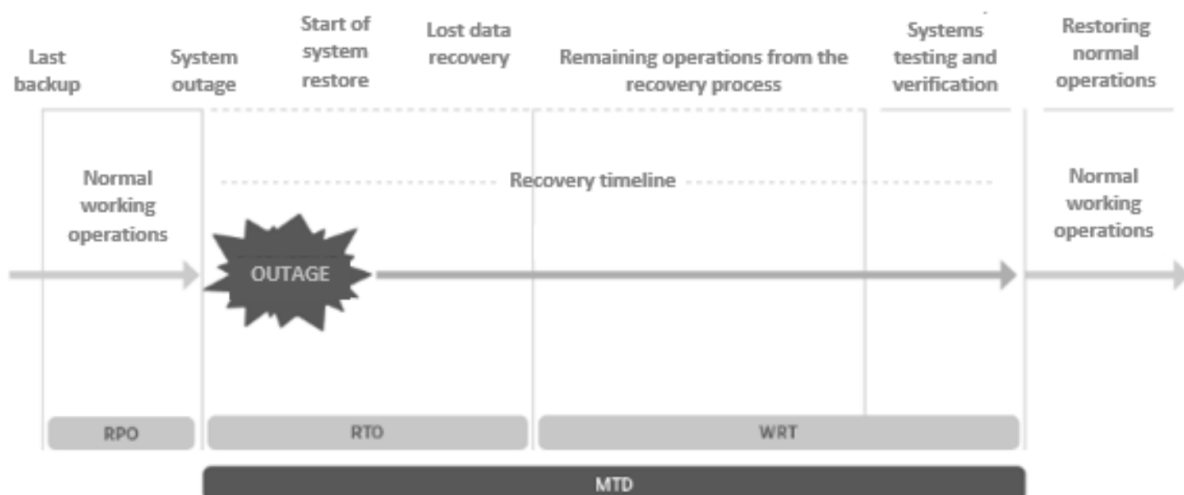


Figure 2: Maximum tolerable downtime in the context of intellectual capital protection

This approach not only ensures operational continuity but also plays a crucial role in protecting the organization's intellectual capital by minimizing downtime and safeguarding essential information systems. In the system design phase, defining requirements for data protection is crucial for ensuring information consistency and safeguarding intellectual capital. A key component in this process is the

RPO, which specifies the maximum age of data in backups at the time of recovery. As a time-dependent metric, the RPO indicates the point to which data will be restored, thus affecting potential data and information loss. This backward recovery process inevitably involves some data loss (except in cases of synchronous replication, where data loss can be eliminated). Consequently, organizations establish a Maximum Tolerable Data Loss (MTDL) threshold during sustainable business continuity planning to ensure data protection within acceptable loss limits. Systems designed to achieve zero data loss reveal an inverse relationship between RPO proximity to the outage and system cost: the closer the RPO, the higher the expense.

Using these parameters as a foundation, our research analyzed the performance and reliability of cloud-based recovery systems following outages. This analysis provides a parametric framework to guide the selection and implementation of recovery systems. By balancing technical, organizational, and financial considerations, this framework assists organizations in making informed choices to protect their intellectual capital, maintain operational continuity, and optimize recovery processes in alignment with business continuity objectives.

#### 4.1. Production environment

This section outlines the working environment for the research, specifically describing the setup of a test virtual machine used to evaluate backup and recovery processes in a real-world, mission-critical environment. Given the need for stability and protection of intellectual capital, a dedicated server was configured as a virtual machine to monitor backup and restoration processes, utilizing cloud-based storage as a Recovery Service. For this purpose, Microsoft Azure was chosen as the cloud service provider, with the Microsoft Azure Recovery Service (MARS) [13] deployed to secure a single virtual machine (VM). Unlike the Disaster Recovery (DR) system discussed in [10], where Backup as a Service (BaaS) is implemented within a data center, our research focuses entirely on a cloud-deployed solution.

The data center supporting the research environment is illustrated in **Figure 3**.

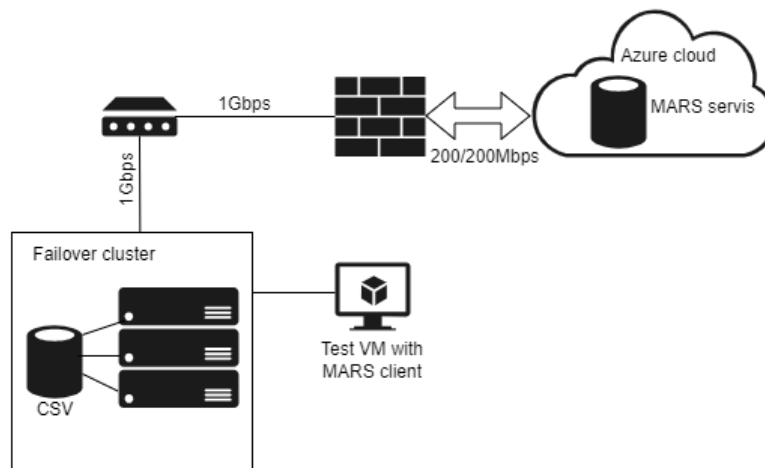


Figure 3: Data Center research environment

This production environment comprises three server systems, each with high fault tolerance across network, data, power, and storage components. The servers, running Windows Server 2019 with Hyper-V, form a virtual cluster that provides a failover mechanism, ensuring high availability (HA) for virtual systems [14]. This cluster-based setup enhances data center availability and service reliability, both locally and across the organization's network.

Due to the data center's strict adherence to operational continuity standards, availability and reliability are considered maximal, consistent with operational uptime goals. Performance success is validated through security and operational logs from the data center systems and Azure services. Microsoft Azure, the cloud provider used in this setup, boasts an impressive availability of nearly

100% (99.99999999% or 11 nines), underscoring its suitability for safeguarding critical information and ensuring uninterrupted access to the organization's intellectual capital.

#### 4.1.1. Parametric prerequisites as a starting point for the research

In defining the key parameters as a starting point for evaluating the solutions in our research, it is essential to recognize that these parameters, from a technical perspective, directly influence the RPO and RTO, two critical metrics for assessing the effectiveness of data protection strategies [12]. Given that the analysis involves dynamic systems where parameter characteristics fluctuate based on operating conditions, it is necessary to consider some parameters as averages. During backup and restore operations, multiple factors, such as data packet delays and fluctuating network traffic, affect these values. The selection of parameters for analysis hinges on the specific objectives of the evaluation, guided by a Business Impact Analysis (BIA). This foundational analysis identifies the crucial parameters for determining the most appropriate recovery solution (primarily RPO and RTO) that align with the organization's needs. By defining these key parameters, organizations can make informed decisions to protect their intellectual capital, ensuring that data recovery strategies are both resilient and responsive to operational demands. In that context, the parameters that are included in creation of a BIA frame, in our research are taken as a basis for the creation of the concept of protection in company's intellectual capital:

- **LAN speed** - 1Gbps
- **Internet connection speed** - 200/200Mbps
- **Backup frequency** - daily
- **RPO** -  $\leq 7$  days
- **RTO** -  $\leq 24$  hours
- **Retention time of backup copies** – 7 days

#### 4.2. System dynamics modeling aproach for the cloud-based system

Figure 4 illustrates the model of the cloud-based system, highlighting four derived components: two related to backup creation, one to data recovery, and one to the monthly cost of using a fully cloud-based service.

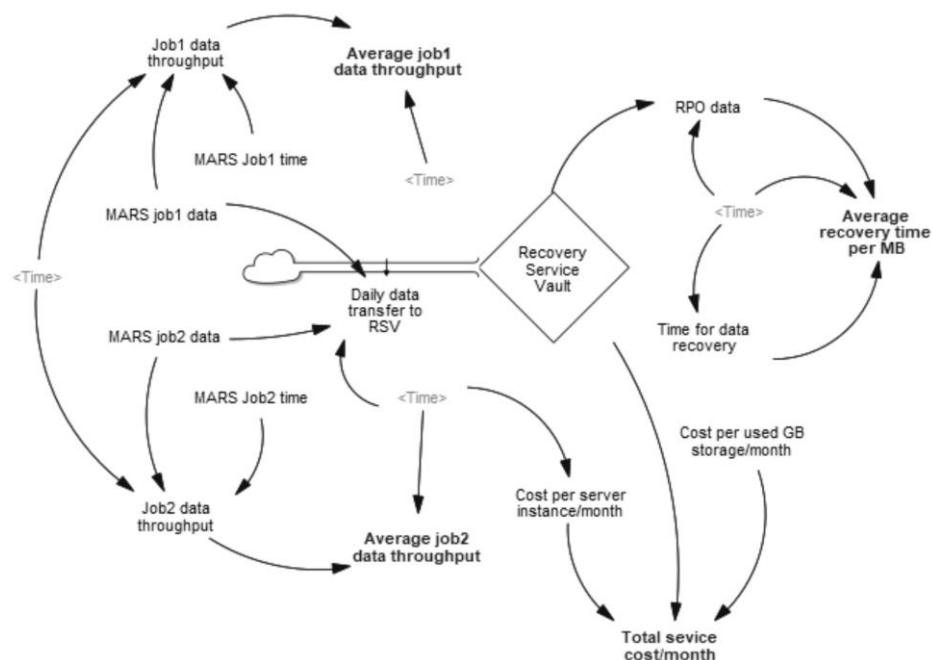


Figure 4: System dynamics model of cloud-based protection system

In the MARS concept, the data backup process involves two key steps, referred to as *job1* and *job2*. In the first step, an agent generates a backup of the virtual machine's system state, capturing essential configuration and operational data. In the second step, a comprehensive backup of remaining data, not covered in the initial step, is performed. This two-step process ensures a thorough safeguard of both system-critical data and broader informational assets, reinforcing the protection of an organization's intellectual capital by securing all facets of the virtual environment. Because of this backup/recovery concept, the model includes **Average job1 data throughput** and **Average job2 data throughput**, which represent the average amount of data transferred to the system's backup storage (Recovery Service Vault - RSV). The variable **Average recovery time per MB** reflects the time needed to recover 1 MB of data from the protective storage located in cloud, while **Total service cost/month** represents the ongoing expense of using cloud-based storage services. This four components are derived from the data captured with MARS agent installed in the protected VM that are referred to the backup/recovery process.

The backup and recovery processes span eight time points. Seven points cover the backup policy cycle, while the final point captures the change in data volume within the recovery vault at the end of a 7-day cycle. Table 1 displays the variable states in the model, showing changes over the specified timeframe, with a separate section summarizing the final values of derived components following the simulation based on the specified input parameters.

By tracking these parameters, organizations can make data-driven decisions on backup and recovery strategies to ensure robust protection of intellectual capital while maintaining cost-efficiency in a cloud environment. The values obtained from the simulation, as presented in Table 1, form the foundation for evaluating system performance under extreme data component conditions. These derived component values directly influence system performance and provide critical insights into how well the system can manage variations. By analyzing these impacts, organizations can better assess the system's capacity to protect intellectual capital, ensuring that backup and recovery processes remain resilient and effective even when subjected to demanding data loads.

**Table 1:**  
Value states of the variables in the model

Variable	Value							
	1	2	3	4	5	6	7	8
<b>Data backup process</b>								
MARS job1 data (MB)	8362	8409	8463	8516	8569	8622	8678	
MARS Job1 time (sec)	3270	2993	3025	3110	2976	3105	3307	
Job1 data throughput (MB/s)	2.55719	2.80956	2.79769	2.73826	2.87937	2.77681	2.62413	
MARS job2 data (MB)	353	375	478	553	551	394	780	
MARS Job2 time (sec)	351	365	489	628	232	358	387	
Job2 data throughput (MB/s)	1.0057	1.0274	0.977505	0.880573	2.375	1.10056	2.0155	
Daily data transfer to RSV (MB)	8715	8784	8941	9069	9120	9016	9458	
<b>Data recovery process</b>								
RPO data (MB)								7690
Time for data recovery (sec)								1380
<b>Derived variables</b>								
Average job1 data throughput (MB/s)								2.57731
Average job2 data throughput (MB/s)								1.83045
Average recovery time per MB (sec)								5.57246
Total service cost/month (dollars)								7.82701

To validate the system's performance in alignment with the requirements emphasized in the BIA, we added five new components to the model. These include two components (**Backup time Job1 (Test data)** and **Backup time Job2 (Test data)**) to calculate the time required for each backup creation step. Another component, **Recovery time (Test data)**, measures the data recovery duration. Additionally, **Test data** serves as a central component representing the total data volume held by the server system within the organization (set at 531 GB). The fifth component, **Total service cost/month (Test data)**, calculates the monthly service cost based on this data volume. These enhancements provide a detailed view of the system's performability, presented in the Table 2:

ensuring that it effectively supports the protection of intellectual capital by accommodating the organization's data recovery and backup needs within cost-effective parameters.

**Table 2:**

Simulation results of the model for a cloud-based system with Test data

Variable	Value
<b>Derived variables</b>	
Average job1 data throughput (MB/s)	2.57731
Average job2 data throughput (MB/s)	1.83045
Average recovery time per MB (sec)	5.57246
<b>Test data simulation values</b>	
Test data (MB)	531012
Backup time Job1 (Test data) (hours)	57.2315
Backup time Job2 (Test data) (hours)	80.5831
Recovery time (Test data) (hours)	26.47
Total service cost/month (Test data) (dollars)	43.7893

If we compare the results obtained from the validation process, with the values requested in the BIA, we will notice that they exceed the maximum allowed in the request of BIA (in the BIA the maximum allowed value for RTO is 24 hours).

## 5. Conclusion

The simulation and research conducted in this study highlight the critical role of cloud-based systems in enhancing business continuity through robust data backup and recovery mechanisms. By thoroughly evaluating the Microsoft Azure Recovery Service - MARS model, we explored how cloud storage solutions can sustain organizational resilience, particularly in scenarios that require immediate data access and rapid recovery. This approach proves essential for safeguarding an organization's intellectual capital, as it ensures continuous access to critical information and assets that form the backbone of organizational knowledge and competitive advantage.

The simulation validates the effectiveness of cloud-based systems in enhancing resilience and protecting intellectual capital. Cloud solutions, exemplified by the MARS model, enable organizations to maintain continuity and safeguard their information assets against disruptions. The insights gained from this study provide a framework for evaluating cloud-based backup and recovery systems, focusing on achieving optimal RTO and RPO, scalability, and cost efficiency. Due to the wide time frame of the cloud-based system in the data recovery processes, such systems are extremely useful and widely applicable for the recovery of systems that do not have critical importance for the business operations of companies or are used as an archive to store a large amount of data or documents for a long period of time. Ultimately, organizations that prioritize intellectual capital protection through robust data management strategies are better positioned to sustain competitive advantage and support long-term growth. The shift towards cloud-based solutions for intellectual capital protection, therefore, represents a strategic imperative in today's data-driven business landscape.

### References:

- [1] A. Mathew, C. Mai, "STUDY OF VARIOUS DATA RECOVERY AND DATA BACK UP TECHNIQUES IN CLOUD COMPUTING & THEIR COMPARISON", 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology, 2018, Bangalore, India
- [2] A.A. Tamimi, R. Dawood, L. Sadaqa, "Disaster Recovery Techniques in Cloud Computing", IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019, Amman, Jordan.

- [3] S. Shahzadi, G. Ubakanma, M. Iqbal, T. Dagiuklas, "Autonomous, Seamless and Resilience Carrier Cloud Brokerage Solution for Business Contingencies During Disaster Recovery", IEEE 20th International Conference on High Performance Computing and Communications, 2018, Exeter, UK.
- [4] T. Wood, E. Cecchet, K. Ramakrishnan, P. Shenoy, J. Van DerMerwe, A. Venkataramani, "Disaster recovery as a cloud service: Economic benefits & deployment challenges", 2nd USENIX Workshop on Hot Topics in Cloud Computing, 2010, Boston.
- [5] Sundać Dragomir, Švast Nataša, "Intelektualni kapital-temeljni čimbenik konkurentnosti poduzeća", Ministarstvo Gospodarstva, Rada i Poduzetništva, Zagreb, 2009, pp. 37.
- [6] A. Mishra, V. Sharma, A. Pandey, "Reliability of Cloud Computing Services", IOSR Journal of Engineering, Volume:04, Issue:01, pp.51-60, 2014.
- [7] A. Mishra, V. Sharma, A. Pandey, "Reliability, Security and Privacy of Data Storage in Cloud Computing", International Journal of Advanced Research in Computer Science, Volume 5, No.3, 2014.
- [8] H.B. Rebah, H.B. Sta, "Disaster Recovery as a Service: A Disaster Recovery Plan in the Cloud for SMEs", Global Summit on Computer & Information Technology, 2016, Sousse, Tunisia.
- [9] J. Mendonça, R. Lima, E. Andradey, J. Araujo, D.S. Kim, "Multiple-criteria Evaluation of Disaster Recovery Strategies Based on Stochastic Models", 16th International Conference on the Design of Reliable Communication Networks (DRCN), IEEE, 2020.
- [10] J. Mendonca, R. Lima, E. Queiroz, E. Andrade, D.S. Kim, "Evaluation of a Backup-as-a-Service Environment for Disaster Recovery", IEEE Symposium on Computers and Communications (ISCC), 2019, Barcelona, Spain.
- [11] J. Mendonca, R. Lima, R. Matos, J. Ferreira, E. Andrade, "Availability Analysis of a Disaster Recovery Solution Through Stochastic Models and fault injection experiments", 32nd International Conference on Advanced Information Networking and Applications, IEEE, 2018.
- [12] S. Nikolovski, P. Mitrevski, "On the Requirements for Successful Business Continuity in the Context of Disaster Recovery", Proc. of the 57th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), Ohrid, North Macedonia, 2022.
- [13] B. Chakraborty, Y. Chowdhury, "Introducing Disaster Recovery with Microsoft Azure: Understanding Services and Tools for Implementing a Recovery Solution", Apress, 2020.
- [14] N. Dhanujati, A.S. Girsang, "Data Center-Disaster Recovery Center (DC-DRC) For High Availability IT Service", International Conference on Information Management and Technology (ICIMTech), 2018, Jakarta.

# Object Tracking in Video Stream

Marko Beljin<sup>1</sup>, Ivan Gasic<sup>1</sup>, Zeljko Stojanov<sup>2</sup>, Peter Odry<sup>3</sup>, Zoltan Vizvari<sup>4</sup> and Vladimir Tadic<sup>1,3,4</sup>

<sup>1</sup> Technical College of Applied Sciences in Zrenjanin, Department of Mechanical Engineering, Electrical Engineering and Computer Science, Zrenjanin, Republic of Serbia

<sup>2</sup> University of Novi Sad, Technical Faculty "Mihajlo Pupin", Department for Information Technologies, Zrenjanin, Republic of Serbia

<sup>3</sup> Institute of Information Technology, University of Dunaujvaros, Tancsics M. Str. 1/A, H-2401 Dunaujvaros, Hungary

<sup>4</sup> John von Neumann Faculty of Informatics, Obuda University, Becsi Str. 96/B, H-1034 Budapest, Hungary

marko.beljin@vts-zr.edu.rs; ivan.gasic@vts-zr.edu.rs; zeljko.stojanov@uns.ac.rs; podry@uniduna.hu; vizvari.zoltan@uni-obuda.hu; \*vladimir.tadic@vts-zr.edu.rs

## Abstract:

This paper focuses on the application of still and moving image processing technology, which plays a key role in modern industrial plants, both in quality control and in the production process itself, through a practical example of a device for object detection and tracking. As a result, an universal video tracker that takes a sample of the captured object from the first frame and analyzes each subsequent frame to find the device in the camera's field of view is obtained. The experiments and results proved that the developed algorithm works well in various object tracking examples.

## Keywords:

Image processing, OpenCV, Object tracking, Embedded Systems

## 1. Introduction

Image processing plays a significant role in modern informatics, with wide range of applications in various industries, including medicine, autonomous systems, surveillance, facial recognition and others. With the increasing accessibility of modern graphic chips that have impressive speeds, image processing technology has become available for various applications, thus the need for efficient and flexible tools for visual data processing is growing. In this context, the OpenCV (Open Source Computer Vision) library stands out as one of the most used solutions for computing that deals with image and video processing.

This library was developed with the aim of making image processing technology available for all users, hence the engineers easily can develop advanced image processing applications. Thanks to its open nature, it has become a global standard for software development in this field. This library offers a wide range of tools for image manipulation, object recognition, edge detection, working with histograms, etc.

Further, this paper includes the application of the mentioned library in practical scenarios, such as motion detection, texture analysis and its use in autonomous systems. A particular advantage of using this library is its compatibility with various programming languages (Python, C++, Java) and its efficiency in processing large data sets. Also, the paper highlights the significance and potentials of OpenCV for further development in the field of image processing, especially in the areas of artificial intelligence and machine vision. The functioning of real-time image and video processing algorithms, as well as the working principle of the device will be explained through a simple machine vision library.

After the introduction, the second section is the related work, which is followed by the theoretical background and the experiments section. In final, the conclusions are given.

## 2. Previous work

Wang et al. [1] introduced objective methods for assessing image quality using perceptual metrics. They tried to estimate how can observers detect visible differences between a reference and a test image, using various known properties of the human visual system. Ojha et al. [2] identified targets in crowded environments. Their work on adapting patients dynamically to video surveillance analysis methods through moving object detection, classification, tracking, and behavior identification is sequentially organized. Their paper considers the following unified issues related to the thematic grouping of such techniques, proposals of such proposed techniques, their types, and the true visualization of critical and efficient group synthesis. Soleimanitaleb et al. [3] primarily focused on learning-based approaches, which are divided into three types: generative models, discriminative models and reinforcement learning. One of the subcategories of the discriminative model is deep learning. Recently, deep learning has received much recognition for its high performance. Comaniciu et al. [4] used a measure-based Bhattacharyya coefficient as a similarity measure and solves optimization using the mean shift procedure. In the presented tracking scenarios, their method managed to overcome problems of camera movement, partial occlusions, clutter and changes in object size. The possibilities of combining it with motion filters and data association methods are also considered. The fields of application are much broader, and only some potential directions are presented such as using background information for OCR, Kalman filters with motion patterns and face tracking. Wang et al. [5] extended the standard mean-shift tracking algorithm to an adaptive tracker that adaptively selects the most reliable colors and textures of shapes according to their ability to describe the target. The objective model is updated based on the similarity index of the original and current model, which increases the reliability of the tracker. The proposed algorithm was evaluated in comparison with some state-of-the-art trackers on demanding image sequences, leading to superior results.

## 3. Object Tracking

This section will introduce the developed object tracking solution. The proposed method uses several well-known image processing algorithms that are included in the OpenCV framework. The object tracking is used to monitor the movement of an object through a series of images (video sequences). The goal is to recognize and track the position of the object over time. First, the object is identified in the initial image using techniques such as:

- Haar cascade
- HOG (Histogram of Oriented Gradients)
- CNN (Convolutional Neural Networks)

Haar cascades are a method for object detection in images, particularly known for their use in facial recognition. It was developed by Paul Viola and Michael Jones in 2001, and it became popular due to its efficiency and speed. These features are similar to those used in Haar functions, which are rectangular patterns in the image. They are used to detect contrasts between different regions of the image, such as light and dark areas, in order to recognize specific patterns. Haar cascades use a series of classifiers in a “cascade” mode. Each classifier in the series focuses on specific features and can quickly discard regions of the image that do not contain the object being searched for.



Figure 1. Haar cascade features



This approach allows for faster processing, as many detection candidates are discarded early in the process, and more complex classifiers are applied only to those parts that have passed through the previous cascades. The cascades are trained using positive and negative examples. Positive examples contain images of the object being recognized (i.e. a face), while negative examples contain images without the object. The algorithm learns to recognize patterns that are characteristic of the object. Thanks to the cascade structure, Haar cascades are extremely fast, making them suitable for real-time applications, such as video surveillance and mobile applications.

HOG analyzes changes in pixel intensity in the image in order to identify edges and shapes. Operators such as the Sobel filter are used to calculate gradients in the horizontal and vertical directions.

The basic operation in CNN involves applying a filter kernel to the input data (i.e. an image) to extract relevant features such as edges, shape textures, etc.

Once the object is detected, several approaches are used for tracking it:

- Kalman filter: Models the movement of the object and predicts its future position [6].
- Mean Shift and CAMShift: These methods focus on changes in the color distribution and shape of the object [7].
- Optical Flow: Analyzes the movement of pixels between consecutive frames to track the movement of the object [8].

All the mentioned methods are implemented in the OpenCV library and can be called with simple commands. The example used for analysis employs a binary image for analysis with the screen display shown in Figure 2 and a grayscale image that is compared in real-time using the aforementioned techniques. By entering commands via the keyboard, pressing the "W", "A", "S", "D" keys moves the red window frame to bring the object into the tracking zone, after which pressing the "T" key activates the object tracking algorithm. As an indication of correct tracking, a blue square will appear around the object that was within the tracking zone. If another object should be tracked, by pressing the "C" key, the new item inside the red square will be framed in blue.

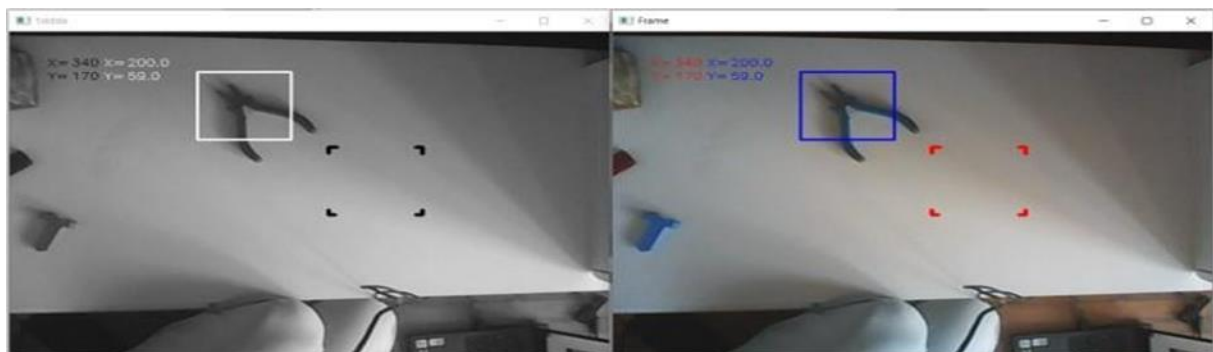


Figure 2. An object tracking example

After the tracking indication, data in the form of coordinates where the square is located will be sent via UART (Universal Asynchronous Receiver-Transmitter) communication. This information is sent to the microcontroller unit (MCU), which uses this data to control the horizontal vertical actuator servo motors (in the following text PAN/TILT), ensuring that the object remains in the center of the camera's field of view. The previously mentioned process is briefly explained in the block diagram in Figure 3.

The PAN/TILT device is designed to allow the camera to move 360° horizontally and 180° vertically, using two DC motors. Such devices are primarily used in video surveillance and FPV (First Person View) drones, where additional feedback from the gyroscope is utilized to enable stabilization of the system while the drone is in motion. In the testing model, the role is played of keeping the object being tracked in the center of the camera's FOV (Field of View).

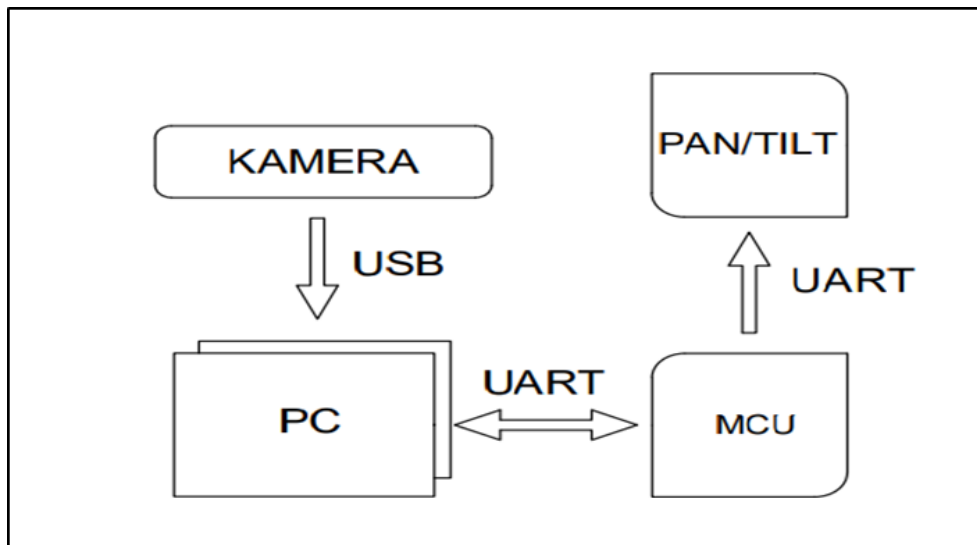


Figure 3. An object tracking example

The software is written in the Python compiler PyCharm and ArduinoIDE for the PAN/TILT software. The mechanical parts are designed in the CAD software CATIA (<https://www.3ds.com/products/catia>), and the parts are manufactured using a 3D printer Ender 3Pro. The control is implemented using a simple H-bridge, which is realized on a prototype board.

#### 4. Experiments and Results

The experiments were conducted using a computer with the following specifications:

- CPU- i5 10500H and 24GB of RAM
- GPU-RTX 3050
- CAMERA-Logitech C170

Utilizing several examples, the software's capability will be tested. Initially, a monochromatic objects with simple geometry will be used. From Figure 4, it can be seen that the object has been successfully tracked, indicated by the blue square around the object. In the upper left corner, the initial and current coordinates of the object are displayed. By further movement of the object, it was established that the tracker is sufficiently capable of tracking simple geometric shapes. It must be emphasized that the surface around the object also plays a significant role.

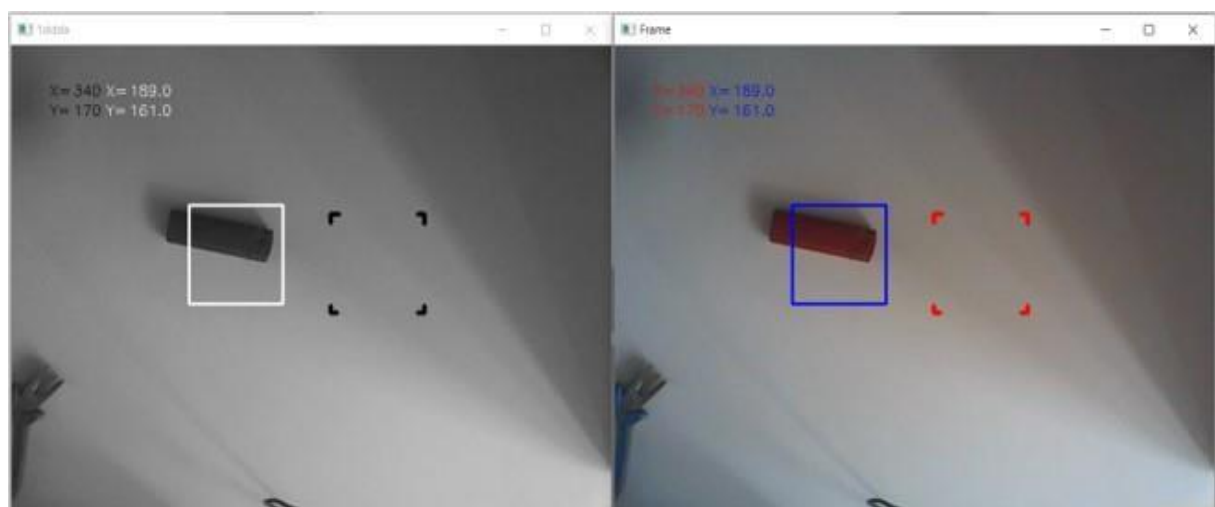


Figure 4. The tracking of the red object

For the second test object, a blue object with a complex geometry was chosen as it is shown in

Figure 5. The object was successfully detected and with further movement it remained in the tracking zone regardless of its geometry and architecture.

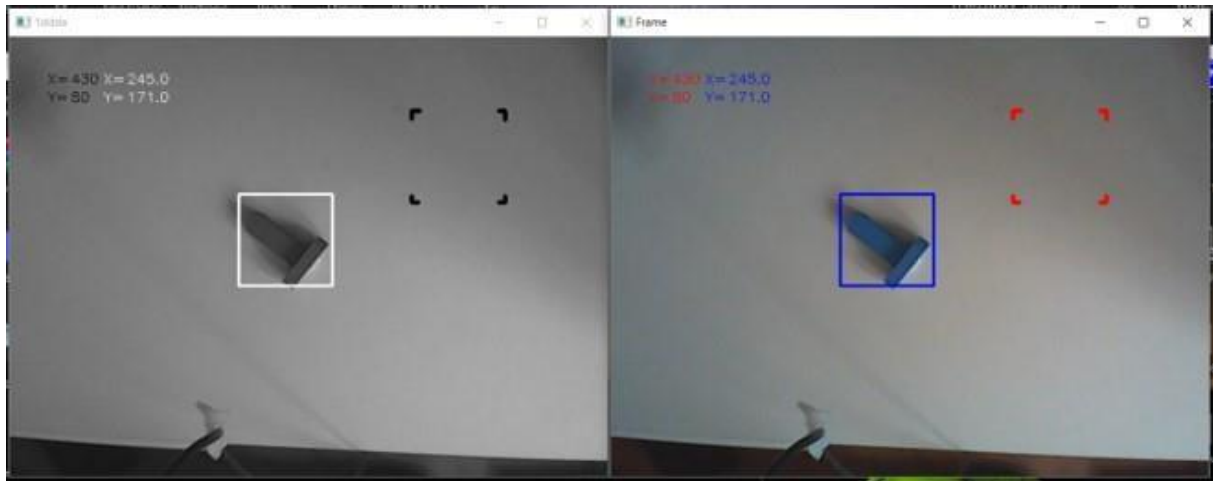


Figure 5. An object tracking example of blue colored complex geometric object

Since monochromatic objects have been observed so far, it was decided that the next object should have a simple geometric shape, but with more colors as it is shown in Figure 6. This object was also appropriately detected, however during the rotation it was lost from the tracking zone. This indicates that this type of tracker has problems with tracking objects that are not monochromatic.

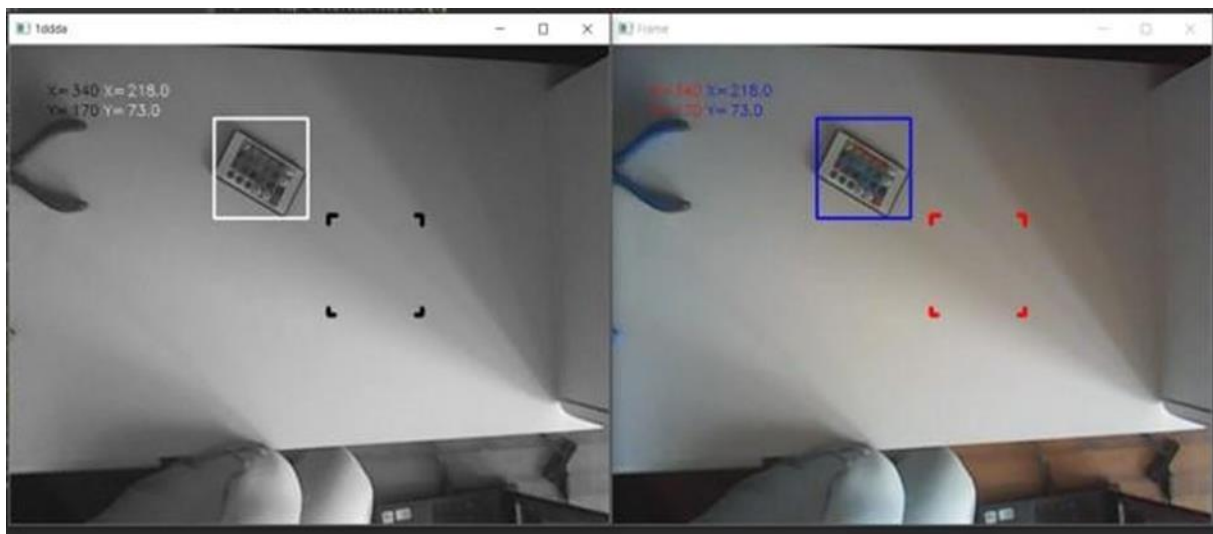


Figure 6. An object tracking example of simple geometric shape with multiple colors

Further improvement of this system will involve the introduction of an adaptive tracker algorithm in order to address the issue of tracking objects that are not monochromatic. Additionally, an algorithm will be implemented to eliminate the problem of object's size variations. The idea is to develop an algorithm that will compare the previous and new positions of the object proportionally, which would increase the reliability of object tracking.

## 5. Conclusion

This paper demonstrates the application of simple algorithms in object detection and tracking. Tracking algorithms use the object's color histogram for tracking. Mean Shift algorithm is used to find the local maxima, while CamShift adaptively updates the size and shape of the search window. Based on the experiments, the proposed algorithm works well. The combination of classical and deep learning

techniques can enhance the resilience to variations and changes in the environment, resulting in more robust tracking systems.

#### **Acknowledgement:**

This research was a part of projects 2020-1.1.2-PIACI-KFI-2020-00173 of the University of Dunaujvaros, and GINOP\_PLUSZ-2.1.1-21-2022-00249 and 2023-1.1.1-PIACI\_FÓKUSZ-2024-00011 of the Obuda University.

#### **References:**

- [1] Z. Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli. (2004). "Image quality assessment: from error visibility to structural similarity", IEEE, pp. 600-612, DOI: 10.1109/TIP.2003.819861.
- [2] S. Ojha and S. Sakhare. (2015). " Image processing techniques for object tracking in video surveillance - A survey", IEEE, 2015 International Conference on Pervasive Computing (ICPC), DOI: 10.1109/PERVASIVE.2015.7087180.
- [3] Z. Soleimanitaleb, M.A. Keyvanrad and A. Jafari. (2019). "Object Tracking Methods:A Review", IEEE, 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), DOI: 10.1109/PERVASIVE.2015.7087180.
- [4] D. Comaniciu, V. Ramesh and P. Meer. (2003). "Kernel-based object tracking", IEEE Transactions on Pattern Analysis and Machine Intelligence ( Volume: 25, Issue: 5, May 2003), pp. 564-577, DOI: 10.1109/TPAMI.2003.1195991.
- [5] J. Wang and Y. Yagi (2008). "Integrating Color and Shape-Texture Features for Adaptive Real-Time Object Tracking", IEEE Transactions on Image Processing ( Volume: 17, Issue: 2, February 2008), pp. 235-240.
- [6] B. Tian, Q. Yao, Y. Gu, K. Wang and Y. Li. (2011). "Video processing techniques for traffic flow monitoring: A survey", 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1103 -1108, DOI: 10.1109/TIP.2007.914150.
- [7] Q. Li, R. Li, K. Ji and W. Dai. (2016). "Kalman Filter and Its Application", IEEE, 2015 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS), DOI: 10.1109/ICINIS.2015.35.
- [8] S. Du. (2023). "CAMShift-Based Moving Object Tracking System", IEEE, 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), DOI: 10.1109/ICETCI57876.2023.10176618.
- [9] Y. Jiao, G. Shi, T. D. Tran. (2021). "Optical Flow Estimation Via Motion Feature Recovery", IEEE, 2021 IEEE International Conference on Image Processing (ICIP), DOI: 10.1109/ICIP42928.2021.9506523
- [10] DAS CATIA <https://www.3ds.com/products/catia>

## Index of Authors

### A

Alavuk, Đorđe	54
Agić, Mugdin	215
Aganović, Mia Lešić	215
Amižić, Vuk	100

### B

Bajrami, Buen	162, 238
Bakator, Mihalj	79, 121, 193
Beljin, Marko	496
Berković, Ivana	310, 396
Bhatt, Madhusudan	280
Blazić, Marko	295, 302
Blazeska Tabakovska, Natasa	181, 247
Bjelica, Jelena	255
Brtka, Eleonora	320, 443, 472, 480
Brtka, Vladimir	262, 420
Bogdanoska Jovanovska, Mimoza	247
Bocevska, Adrijana	17, 93, 106
Bodiroga, Dragana	85
Boral, Edit	396
Bychkov, Igor V.	28
Bursać, Vasilije	173
Bulajic, Aleksandar	45
Burić, Dragan	272

### C

Chotaliya, Narendra	280
---------------------	-----

### D

Dašić, Predrag	326
Denić, Nebojša	402
Desnica, Eleonora	436
Divitarova, Evgenija	93
Dimitriovski, Ivica	451
Dinu, Dragan	132, 148
Dobrilović, Dalibor	64, 355, 436
Dobardžić, Dilan	262, 267, 295
Dragović, Dragan	64, 355, 363
Dragović, Rade	363
Drašković, Milana	156

### Đ

Đuričić, Milenko	355
Đurđev, Mića	193
Đorđević, Luka	79, 193
Đukić, Dejan	223, 231, 402
Đukić Popović, Sonja	402

### E

Eremić, Željko	156, 341
----------------	----------

### F

Fereferov, Evgeniy S.	28
Fedorov, Roman K.	28

### G

Gaborov, Maja	200, 334
Gajić B., Dušan	148
Gluvakov, Verica	114, 121
Glušac, Dragana	320, 408
Gašić, Ivan	496

### H

Habibulla, Gisele	208
Hristoski, Ilija	140

### I

Ilić, Velibor	35
Ivetić, Dragan	72, 85, 173

### J

Jakovljević, Ana	272
Jauševac, Goran	200, 472
Jančev, Nikola	54
Jerković, Nikola	310, 396
Jokić, Snežana	54
Jovanović, Ana	396
Jolevski, Ilija	162
Jotanović, Gordana	200, 472
Jovanović, Mitar	114
Jovanov, Nikola	320, 443, 480

### K

Kačanovski, Vladimir	363
Kazi, Ljubica	280
Kazi, Zoltan	100, 280
Kavalić, Mila	114, 121
Kupusinac, Aleksandar	72, 173, 316
Knežević, Maša	187
Kovačević, Miodrag	334, 408
Kovač, Dragana	200, 334, 420
Kószó, Albina	480
Kotevski, Zoran	17, 106
Kreculj, Dragan	334
Krneta, Dragoljub	380

### L

Lazić, Olivera	272
Lazić Vulićević, Ljubica	341

### Lj

Ljubojev, Nadežda	64, 355,
-------------------	----------

### M

Markoski, Aleksandar	17, 106, 287
Markoski, Branko	187
Manevska, Violeta	238
Mazalica, Milica	443
Mladenović, Valentina	341
Milošev Vladimir	64
Mirković Snežana	114, 121
Milić. Tamara	255, 408
Matkovski, Marijana	272

Marković, Milan	295, 302	<b>S</b>	
Makitan, Vesna	320, 408, 480	Savoska, Snežana	247
Milenkovski, Božidar	488	Slavkovska, Daniela	17, 451, 461, 488
<b>N</b>		Simić, Ivana	127
Naneva, Veselina	347	Simić, Jelena	472
Nadrljanski, Slobodan	302	Stefanova, Kremena	347
Nedelkovski, Igor	17, 106	Skejić, Emir	215
Nikolić, Boško	372	Siljanoska, Teodora	181
Nikolovski, Saso	461, 488	Stanisavljev, Sanja	114, 121
Ninkov, Magdalena	54	Sladić, Goran	187
Novokmet, Predrag	320, 443, 480	Stojanov, Jelena	140, 262, 420
Novaković, Milada	64	Stojanov, Željko	200, 496
Novaković, Borivoj	79, 193	Stojković, Lucija	127
Nurkanovic, Ajla	472	Stupar, Olivera	114
<b>O</b>		Stojkov Lončarski, Aleksandra	187, 443
Odry, Peter	496	Shilakadze, Vazha	326
O'Connor, Slavica	2	Semanová, Katarína	388
Ognjenović, Višnja	262, 302, 310, 372	Samardžić, Ivana	380
Ognjenović, Ana Marija	372	Szabó, Csaba	388
<b>P</b>		<b>Š</b>	
Palinkaš, Ivan	436	Šarenac, Uroš	295
Pardanjac, Marjana	54	Šeljmeši, Dalibor	420
Pekez, Jasmina	436	<b>T</b>	
Peraković, Dragan	420	Tasić, Nemanja	334, 408
Petrović B., Veljko	148	Tadić, Vladimir	496
Petreska, Marija	247	Terek Stojanović, Edit	413
Petreska, Anita	451, 461, 488	Torbica, Bojan	72
Perović, Bojan	355, 363	Trajkovska, Aneta	287, 430
Premčevski, Velibor	79	Turmanidze, Raul	326
Poparić, Sara	148	<b>U</b>	
Popov, Srđan	132, 200, 302	Ugrinov, Stefan	79, 121, 193, 413
Popović, Stefan	231, 402	<b>V</b>	
Popkhadze, Giorgi	326	Veljanovska, Kostandina	17, 162, 238, 430
Prodanović, Armin	472	Veljanovski, Nikolce	430
<b>R</b>		Vecštejn, Igor	200
Radovanović, Zlatko	223	Vignjević, Katarina	262, 295, 302
Rendevski, Nikola	461	Vizvari, Zoltan	496
Radulović, Biljana	295	Vranić, Tanja	132
Ristevski, Blagoj	17, 451, 461	Vrbaški, Dunja	272
Ratković Kovačević, Nada	334	<b>Z</b>	
		Zeljkić, Milena	132

## ***Company Presentation***

### **Advanced Cyber Security (ACS)**

**Advanced Cyber Security** is a company founded in 2020 by experts with extensive international experience in the field of cyber security, including former employees of the Republic of Serbia's Government and private companies specializing in information security. We offer years of experience in practicing offensive and defensive cyber security within the military, governmental, and private sectors.

At this year's **Adria Security Summit, held on October 9 and 10 in Sarajevo**, Advanced Cyber Security (ACS) **was a Platinum sponsor and winner of the Adria Excellence Award**.

This prestigious regional event brought together leading experts in the field of security and technology, and ACS took home the **Adria Excellence Award for innovation in the cybersecurity domain thanks to its ARMADA AI Engine Module**.

<https://www.asadria.com/acs-armada-stiti-prestiznu-banku-od-cyber-napada/>